Note: Before using this information and the product it supports, read the general information under Notices on page 213.

This edition applies to IBM® SPSS® Exact Tests 22 and to all subsequent releases and modifications until otherwise indicated in new editions.

Microsoft product screenshots reproduced with permission from Microsoft Corporation.

We owe a special debt to Professor Marvin Zelen for creating an exciting intellectual environment in the Department of Biostatistics at Harvard. He encouraged us to work on a number of challenging research problems in computational statistics, and this research has culminated in the development of Exact Tests.

Cyrus R. Mehta and Nitin R. Patel
Cytel Software Corporation and Harvard School of Public Health
Cambridge, Massachusetts

Contents

1 Getting Started

The Exact Tests option provides two new methods for calculating significance levels for the statistics available through the Crosstabs and Nonparametric Tests procedures. These new methods, the exact and Monte Carlo methods, provide a powerful means for obtaining accurate results when your data set is small, your tables are sparse or unbalanced, the data are not normally distributed, or the data

Figure 1.1 shows results from an entrance examination for fire fighters in a small township. This data set compares the exam results based on the race of the applicant.

Figure 1.1 Fire fighter entrance exam results

Test Results * Race of Applicant Crosstabulation

Count

		Race of Applicant			
		White	Black	Asian	Hispanic
Test Results	Pass	5	2	2	
	No Show		1		1
	Fail		2	3	4

The data show that all five white applicants received a *Pass* result, whereas the results for the other groups are mixed. Based on this, you might want to test the hypothesis that exam results are not independent of race. To test this hypothesis, you can run the Pearson chi-square test of independence, which is available from the Crosstabs procedure. The results are shown in Figure 1.2.

Figure 1.2 Pearson chi-square test results for fire fighter data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	11.556[1]	6	.073

1. 12 cells (100.0%) have expected count less than 5. The minimum expected count is .50.

Because the observed significance of 0.073 is larger than 0.05, you might conclude that exam results are independent of race of examinee. However, notice that the data contains only twenty observations, that the minimum expected frequency is 0.5, and that all 12 of the cells have an expected frequency of less than 5. These are all indications that the assumptions necessary for the standard asymptotic calculation of the significance level

for this test may not have been met. Therefore, you should obtain exact results. The exact results are shown in Figure 1.3.

The exact *p* value based on Pearson's statistic is 0.040, compared to 0.073 for the asymptotic value. Using the exact *p* value, the null hypothesis would be rejected at the 0.05 significance level, and you would conclude that there is evidence that the exam results and race of examinee are related. This is the opposite of the conclusion that would have been reached with the asymptotic approach. This demonstrates that when the assumptions of the asymptotic method cannot be met, the results can be unreliable. The exact calculation always produces a reliable result, regardless of the size, distribution, sparseness, or balance of the data.

The Monte Carlo Method

Although exact results are always reliable, some data sets are too large for the exact *p* value to be calculated, yet don't meet the assumptions necessary for the asymptotic method. In this situation, the Monte Carlo method provides an unbiased estimate of the exact *p* value, without the requirements of the asymptotic method. (See Ta735 0877.0003 n3asuT

ified number of these possible tables in order to obtain an unbiased estimate of the true *p* value. Figure 1.4 displays the Monte Carlo results for the fire fighter data.

Figure 1.4 Monte Carlo results of the Pearson chi-square test for fire fighter data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Monte Carlo Significance (2-tailed)		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Pearson Chi-Square	11.556[1]	6	.073	.041[2]	.036	.046

1. 12 cells (100.0%) have expected count less than 5. The minimum expected count is .50.
2. Based on 10000 and seed 2000000 ...

The Monte Carlo estimate of the *p* value is 0.041. This estimate was based on 10,000 samples. Recall that the exact *p* value was 0.040, while the asymptotic *p* value is 0.073. Notice that the Monte Carlo estimate is extremely close to the exact value. This demonstrates that if an exact *p* value cannot be calculated, the Monte Carlo method produces an unbiased estimate that is reliable, even in circumstances where the asymptotic *p* value is not.

When to Use Exact Tests

Calculating exact results can be computationally intensive, time-consuming, and can sometimes exceed the memory limits of your machine. In general, exact tests can be performed quickly with sample sizes of less than 30. Table 1.1 and Table 1.2 provide a guideline for the conditions under which exact results can be obtained quickly. In Table 1.2, *r* indicates rows, and *c* indicates columns in a contingency table.

Table 1.1 Sample sizes (N) at which the exact p values for nonparametric tests are computed quickly

One-sample inference	
Chi-square goodness-of-fit test	$N \leq 30$
Binomial test and confidence interval	$N \leq 100,000$
Runs test	$N \leq 20$
One-sample Kolmogorov-Smirnov test	$N \leq 30$
Two-related-sample inference	
Sign test	$N \leq 50$
Wilcoxon signed-rank test	$N \leq 50$
McNemar test	$N \leq 100,000$
Marginal homogeneity test	$N \leq 50$
Two-independent-sample inference	
Mann-Whitney test	$N \leq 30$
Kolmogorov-Smirnov test	$N \leq 30$
Wald-Wolfowitz runs test	$N \leq 30$
K-related-sample inference	
Friedman's test	$N \leq 30$
Kendall's *W*	$N \leq 30$
Cochran's *Q* test	$N \leq 30$
K-independent-sample inference	
Median test	$N \leq 50$
Kruskal-Wallis test	$N \leq 15, K \leq 4$
Jonckheere-Terpstra test	$N \leq 20, K \leq 4$
Two-sample median test	$N \leq 100,000$

Table 1.2 Sample sizes (N) and table dimensions (r, c) at which the exact p values for Crosstabs tests are computed quickly

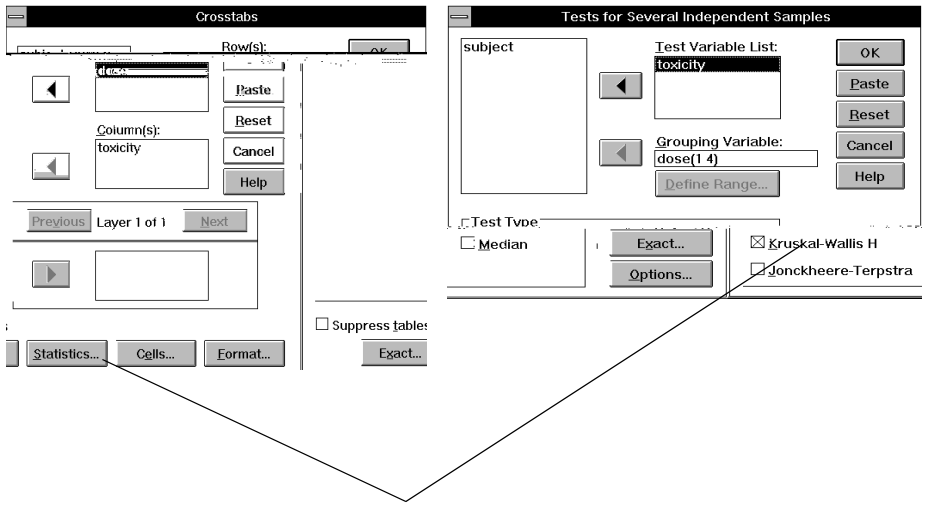
2 x 2 contingency tables (obtained by selecting chi-square)	
Pearson chi-square test	$N \leq 100,000$
Fisher's exact test	$N \leq 100,000$
Likelihood-ratio test	$N \leq 100,000$
r x c contingency tables (obtained by selecting chi-square)	
Pearson chi-square test	$N \leq 30$ and $\min(r, c) \leq 3$
Fisher's exact test	$N \leq 30$ and $\min(r, c) \leq 3$
Likelihood-ratio test	$N \leq 30$ and $\min(r, c) \leq 3$
Linear-by-linear association test	$N \leq 30$ and $\min(r, c) \leq 3$
Correlations	
Pearson's product-moment correlation coefficient	$N \leq 7$
Spearman's rank-order correlation coefficient	$N \leq 10$
Ordinal data	
Kendall's tau-*b*	$N \leq 20$ and $r \leq 3$
Kendall's tau-*c*	$N \leq 20$ and $r \leq 3$
Somers' *d*	$N \leq 30$
Gamma	$N \leq 20$ and $r \leq 3$
Nominal data	
Contingency coefficients	$N \leq 30$ and $\min(r, c) \leq 3$
Phi and Cramér's *V*	$N \leq 30$ and $\min(r, c) \leq 3$
Goodman and Kruskal's tau	$N \leq 20$ and $r \leq 3$
Uncertainty coefficient	$N \leq 30$ and $\min(r, c) \leq 3$
Kappa	$N \leq 30$ and $c \leq 5$

How to Obtain Exact Statistics

The exact and Monte Carlo methods are available for Crosstabs and all of the Nonparametric tests.

To obtain exact statistics, open the Crosstabs dialog box or any of the Nonparametric Tests dialog boxes. The Crosstabs and Tests for Several Independent Samples dialog boxes are shown in Figure 1.5.

Figure 1.5 Crosstabs and Nonparametric Tests dialog boxes



Select the statistics that you want to calculate. To select statistics in the Crosstabs dialog box, click Statistics.

To select the exact or Monte Carlo method for computing the significance level of the selected statistics, click Exact in the Crosstabs or Nonparametric Tests dialog box. This opens the Exact Tests dialog box, as shown in Figure 1.6.

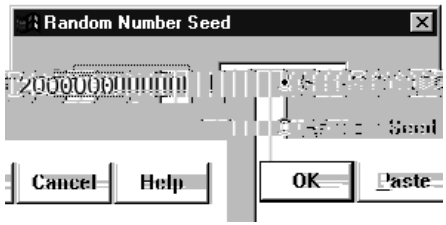
You can choose one of the following methods for computing statistics. The method you choose will be used for all selected statistics.

Asymptotic only. Calculates significance levels using the asymptotic method. This provides the same results that would be provided without the Exact Tests option.

Monte Carlo. Provides an unbiased estimate of the exact *p* value and displays a confidence interval using the Monte Carlo sampling method. Asymptotic results are also displayed. The Monte Carlo method is less computationally intensive than the exact method, so results can often be obtained more quickly. However, if you have chosen the Monte Carlo method, but exact results can be

want to repeat an analysis. To reset the seed, open the Random Number Seed dialog box from the Transform menu. The Random Number Seed dialog box is shown in Figure 1.7.

Figure 1.7 Random Number Seed dialog box



Set seed to. Specify any positive integer value up to 999,999,999 as the seed value. The seed is reset to the specified value each time you open the dialog box and click on OK. The default seed value is 2,000,000.

To duplicate the same series of random numbers, you should set the seed *before* you generate the series for the first time.

Random seed. Sets the seed to a random value chosen by your system.

Pivot Table Output

With this release of Exact Tests, output appears in pivot tables. Many of the tables shown in this manual have been edited by pivoting them, by hiding categories that are not relevant to the current discussion, and to show more decimal places than appear by default.

generated from multinomial, hypergeometric, or Poisson distributions is chi-square. This work was found to be applicable to a whole class of discrete data problems. It was followed by significant contributions by, among others, Yule (1912), R. A. Fisher (1925, 1935), Yates (1984), Cochran (1936, 1954), Kendall and Stuart (1979), and Goodman (1968) and eventually evolved into the field of categorical data analysis. An excellent up-to-date textbook dealing with this rapidly growing field is Agresti (1990).

The techniques of nonparametric and categorical data inference are popular mainly because they make only minimal assumptions about how the data were generated—assumptions such as independent sampling or randomized treatment assignment. For continuous data, you do not have to know the underlying distribution giving rise to the data. For categorical data, mathematical models like the multinomial, Poisson, or hypergeometric model arise naturally from the independence assumptions of the sampled observations. Nevertheless, for both the continuous and categorical cases, these methods do require one assumption that is sometimes hard to verify. They assume that the data set is large enough for the test statistic to converge to an appropriate limiting normal or chi-square distribution. *P* values are then obtained by evaluating the tail area of the limiting distribution, instead of actually deriving the true distribution of the test statistic and then evaluating its tail area. *P* values based on the large-sample assumption are known as *asymptotic p* values, while *p* values based on deriving the true distribution of the test statistic are termed *exact p* values. While exact *p* values are preferred for scientific inference, they often pose formidable computational problems and so, as a practical matter, asymptotic *p* values are used in their place. For large and well-balanced data sets, this makes very little difference, since the exact and asymptotic *p* values are very similar. But for small, sparse, unbalanced, and heavily tied data, the exact and asymptotic *p* values can be quite different and may lead to opposite conclusions concerning the hypothesis of interest. This was a major concern of R. A. Fisher, who stated in the preface to the first edition of *Statistical Methods for Research Workers* (1925):

> The traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small problems on their merits does it seem possible to apply accurate tests to practical data.

The exact p value obtained above is 0.001, implying that there is a strong row and column interaction. Chapter 9 discusses this and related tests in detail.

The above example highlights the need to compute the exact p value, rather than relying on asymptotic results, whenever the data set is small, sparse, unbalanced, or heavily tied. The trouble is that it is difficult to identify, a priori, that a given data set suffers from these obstacles to asymptotic inference. Bishop, Fienberg, and Holland (1975), express the predicament in the following way.

The difficulty of exact calculations couple

Pearson Chi-Square Test for a 3 x 4 Table

Figure 2.4 shows results from an entrance examination for fire fighters in a small township.

Figure 2.4 Fire fighter entrance exam results

Test Results * Race of Applicant Crosstabulation

Count

		Race of Applicant			
		White	Black	Asian	Hispanic
Test Results	Pass	5	2	2	
	No Show		1		1
	Fail		2	3	4

The table shows that all five white applicants received a *Pass* result, whereas the results for the other groups are mixed. Is this evidence that entrance exam results are related to race? Note that while there is some evidence of a pattern, the total number of observations is only twenty. Null and alternative hypotheses might be formulated for these data as follows:

Null Hypothesis: Exam results and race of examinee are independent.

ndence,
test was

ude that
at table
reports that the minimum expected frequency is 0.5, and that all 12 of the cells have an expected frequency that is less than five.

That is, the application warns you that all of the cells in the table have small expected counts. What does this mean? Does it matter?

Recall that the Pearson chi-square statistic, , is computed from the observed and the expected counts under the null hypothesis of independence as follows:

Equation 2.1

you can repeat this analysis for every single table in the reference set, identify all those that are at least as extreme as the original table, and sum their exact hypergeometric probabilities. The exact *p* value is this sum.

Exact Tests produces the following result:

$$\Pr(X^2 \geq 11.55556) = 0.0398 \qquad \text{Equation 2.5}$$

The exact results are shown in Figure 2.6.

Pearson Chi-Square	11.556[1]	6	.073	.040

The exact *p* value based on Pearson's statistic is 0.040. At the 0.05 level of significance, the null hypothesis would be rejected and you would conclude that there is evidence that the exam results and race of examinee are related. This conclusion is the opposite of the conclusion that would be reached with the asymptotic approach, since the latter produced a *p* value of 0.073. The asymptotic *p* value is only an approximate estimate of the exact *p* value. Kendall and Stuart (1979) have proved that as the sample size goes to infinity, the exact *p* value (see Equation 2.5) converges to the chi-square based *p* value (see Equation 2.3). Of course, the sample size for the current data set is not infinite, and you can observe that this asymptotic result has fared rather poorly.

Fisher's Exact Test for a 2 x 2 Table

It could be said that Sir R. A. Fisher was the father of exact tests. He developed what is popularly known as Fisher's exact test for a single contingency table. His motivating example was as follows (see Agresti, 1990, for a related discussion). When drinking tea, a British woman claimed to be able to distinguish whether milk or tea was added to the cup first. In order to test this claim, she was given eight cups of tea. In four of the cups, tea was added first, and in four of the cups, milk was added first. The order in which the cups were presented to her was randomized. She was told that there were four cups of each type, so that she should make four predictions of each order. The results of the experiment are shown in Figure 2.7.

The Pearson chi-square test of independence can be calculated to test this hypothesis. This example tests the alternative hypothesis at the 0.05 significance level. Results are shown in Figure 2.8.

The reported significance, 0.157, is two-sided. Because the alternative hypothesis is one-sided, you might halve the reported significance, thereby obtaining 0.079 as the observed *p* value. Because the observed *p* value is greater than 0.05, you might conclude that there is no evidence that the woman can correctly guess tea-milk order, although the observed level of 0.079 is only marginally larger than the 0.05 level of significance used for the test.

It is easy to see from inspection of Figure 2.7 that the expected cell count under the null hypothesis of independence is 2 for every cell. Given the popular rules of thumb about expected cell counts cited above, this raises concern about use of the one-degree-of-freedom chi-square distribution as an approximation to the distribution of the Pearson chi-square statistic for the above table. Rather than rely on an approximation that has an asymptotic justification, suppose you can instead use an exact approach.

For the table, Fisher noted that under the null hypothesis of independence, if you assume fixed marginal frequencies for both the row and column marginals, then the hypergeometric distribution characterizes the distribution of the four cell counts in the table. This fact enables you to calculate an exact *p* value rather than rely on an asymptotic justification.

Let the generic four-fold table, , take the form

with being the four cell counts; and , the row totals; and , the column totals; and , the table total. If you assume the marginal totals as given,

$$\Pr(x_{ij}) = \frac{\binom{m_1}{x_{11}}\binom{m_2}{n_1 - x_{11}}}{\binom{N}{n_1}} \qquad \text{Equation 2.6}$$

The p value for Fisher's exact test of independence in the 2 2 table is the sum of hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcome.

Let's apply this line of thought to the tea drinking problem. In this example, the experimental design itself fixes both marginal distributions, since the woman was asked to guess which four cups had the milk added first and therefore which four cups had the tea added first. So, the table has the following general form:

Guess	**Pour**		**Row Total**
	Milk	**Tea**	
Milk	x_{11}	x_{12}	4
Tea	x_{21}	x_{22}	4
Col_Total	**4**	**4**	**8**

Focusing on x_{11}, this cell count can take the values 0, 1, 2, 3, or 4, and designating a value for x_{11} determines the other three cell values, given that the marginals are fixed. In other words, assuming fixed marginals, you could observe the following tables with the indicated probabilities:

	Table			**Pr(Table)**	**value**
$x_{11} = 0$	0	4	4	0.014	1.000
	4	0	4		
	4	4	8		
$x_{11} = 1$	1	3	4	0.229	0.986
	3	1	4		
	4	4	8		
$x_{11} = 2$	2	2	4	0.514	0.757

observed contingency table are naturally fixed is irrelevant to the method used to compute the exact test. In either case, you compute an exact *p* value by examining the observed table in relation to all other tables in a reference set of contingency tables whose margins are the same as those of the actually observed table. You will see that the idea behind this relatively simple example generalizes to include all of the nonparametric and categorical data settings covered by Exact Tests.

Choosing between Exact, Monte Carlo, and Asymptotic P Values

The above examples illustrate that in order to compute an exact *p* value, you must enumerate all of the outcomes that could occur in some reference set besides the outcome that was actually observed. Then you order these outcomes by some measure of discrepancy that reflects deviation from the null hypothesis. The exact *p* value is the sum of exact probabilities of those outcomes in the reference set that are at least as extreme as the one actually observed.

Enumeration of all of the tables in a reference set can be computationally intensive. For example, the reference set of all 5 6 tables of the form

The method has the additional advantage that it takes a predictable amount of time, and an answer is available at any desired level of accuracy.

Exact Tests makes it very easy to move back and forth between the exact and Monte Carlo options. So feel free to experiment.

The following sections discuss the exact, Monte Carlo, and asymptotic *p* values in greater detail.

When to Use Exact P Values

Ideally you would use exact *p* values all of the time. They are, after all, the gold standard. Only by deciding to accept or reject the null hypothesis on the basis of an exact *p* value are you guaranteed to be protected from type 1 errors at the desired significance level. In practice, however, it is not possible to use exact *p* values all of the time. The algorithms in Exact Tests might break down as the size of the data set increases. It is difficult to quantify just how large a data set can be solved by the exact algorithms, because that depends on so many factors other than just the sample size. You can sometimes compute an exact *p* value for a data set whose sample size is over 20,000, and at othere size of the

Figure 2.10 Left ventricular wall thickness versus sports activity

Count

		Left Ventricular Wall Thickness		
		>= 13 mm	< 13 mm	Total
SPORT	Weightlifting	1	6	7
	Field wt. events		9	9
	Wrestling/Judo		16	16
	Tae kwon do	1	16	17
	Roller Hockey	1	22	23
	Team Handball	1	25	26
	Cross-coun. skiing	1	30	31
	Alpine Skiing		32	32
	Pentathlon		50	50
	Roller Skating		58	58
	Equestrianism		28	28
	Bobsledding	1	15	16
	Volleyball		51	51
	Diving	1	10	11
	Boxing		14	14
	Cycling	1	63	64
	Water Polo		21	21
	Yatching		24	24
	Canoeing	3	57	60
	Fencing	1	41	42
	Tennis		47	47
	Rowing	4	91	95
	Swimming		54	54
	Soccer		62	62
	Track		89	89

You can obtain the results of the likelihood-ratio statistic for this 25 2 contingency table with the Crosstabs procedure. The results are shown in Figure 2.11.

Figure 2.11 Likelihood ratio for left ventricular wall thickness versus sports activity data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Likelihood Ratio	32.495	24	.115

The value of this statistic is 32.495. The asymptotic *p* value, based on the likelihood-ratio test, is therefore the tail area to the right of 32.495 from a chi-square distribution with 24 degrees of freedom. The reported *p* value is 0.115. But notice how sparse and unbalanced this table is. This suggests that you ought not to rely on the asymptotic *p* value. Ideally, you would like to enumerate every single 25 2 contingency table with the same row and column margins as those in Figure 2.10, identify tables that are more extreme than the observed table under the null hypothesis, and thereby obtain the exact *p* value. This is a job for Exact Tests. However, when you try to obtain the exact likelihood-ratio *p* value in this manner, Exact Tests gives the message that the problem is too large for the exact option. Therefore, the next step is to use the Monte Carlo option. The Monte Carlo option can generate an extremely accurate estimate of the exact *p* value by sampling 25 2 tables from the reference set of all tables with the observed margins a large number of times. The default is 10,000 times, but this can easily be changed in the dialog box. Provided each table is sampled in proportion to its hypergeometric probability (see Equation 2.4), the fraction of sampled tables that are at least as extreme as the observed table gives an unbiased estimate of the exact *p* value. That is, if M tables are sampled from the reference set, and Q of them are at least as extreme as the observed table (in the sense of having a likelihood-ratio statistic greater than or equal to 32.495), the Monte Carlo estimate of the exact *p* value is

$$\hat{p} = \frac{Q}{M} \qquad \text{Equation 2.7}$$

The variance of this estimate is obtained by straightforward binomial theory to be:

$$\text{var}(\hat{p}) = \frac{p(1-p)}{M} \qquad \text{Equation 2.8}$$

Thus, a % confidence interval for *p* is

Equation 2.9

where is the th percentile of the standard normal distribution. For example, if you wanted a 99% confidence interval for *p*, you would use . This is the default in Exact Tests, but it can be changed in the dialog box. The Monte Carlo results for these data are shown in Figure 2.12.

The Monte Carlo estimate of 0.044 for the exact *p* value is based on 10,000 random samples from the reference set, using a starting seed of 2000000. Exact Tests also computes a 99% confidence interval for the exact *p* value. This confidence interval is (0.039, 0.050). You can be 99% sure that the true *p* value is within this interval. The width can be narrowed even further by sampling more tables from the reference set. That will reduce the variance (see tterval 36oval 30.059dbf his cow[(will re3.5tion 488059d)-[(d)-rcentilue. Thi

interval (see Equation 2.9). It is a simple matter to sample 50,000 times from the reference set instead of only 10,000 times. These results are shown in Figure 2.13.

Figure 2.13 Monte Carlo results with sample size of 50,000

Likelihood Ratio	32.495	24	.115	.045[2]	.043	.047

With a sample of size 50,000 and the same starting seed, 2000000, you obtain 0.045 as the Monte Carlo estimate of *p*. Now the 99% confidence interval for *p* is (0.043, 0.047).

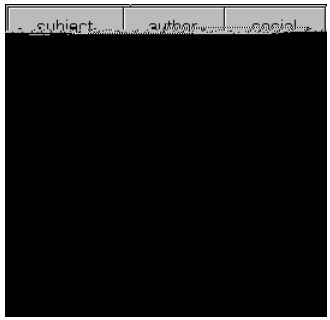
How good are the Monte Carlo estimates? Why would you use them rather than the asymptotic *p* value of 0.115? There are several major advantages to using the Monte Carlo method as opposed to using the asymptotic *p* value for inference.

1. The Monte Carlo estimate is unbiased. That is, .
2. The Monte Carlo estimate is accompanied by a confidence interval within which the exact *p* value is guaranteed to lie at the specified confidence level. The asymptotic *p* value is not accompanied by any such probabilistic guarantee.
3. The width of the confidence interval can be made arbitrarily small, by sampling more tables from the reference set.
4. In principle, you could narrow the width of the confidence interval to such an extent that the Monte Carlo *p* value becomes indistinguishable from the exact *p* value up to say the first three decimal places. For all practical purposes, you could then claim to have the exact *p* value. Of course, this might take a few hours to accomplish.
5. In practice, you don't need to go quite so far. Simply knowing that the upper bound of the confidence interval is below 0.05, or that the lower b

The result, based on 10,000 observations and a starting seed of 2000000, is 0.041. This is much closer to the exact *p* value for the Pearson test, 0.040, than the asymptotic *p* value, 0.073. As an exercise, run the Monte Carlo version of the Pearson test on this data set a few times with different starting seeds. You will observe that the Monte Carlo estimate changes

Small Data Sets. The data set shown in Figure 2.15 consists of the first 7 pairs of observations of the authoritarianism versus social status striving data discussed in Siegel and Castellan (1988).

Figure 2.15 Subset of authoritarianism versus social status striving data



Pearson's product-moment correlation coefficient computed from this sample is 0.7388. This result is shown in Figure 2.16.

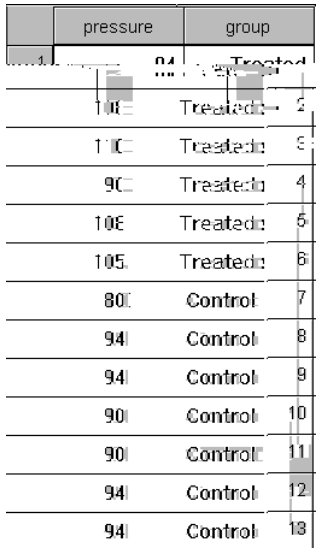
Figure 2.16 Pearson's product-moment correlation coefficient for social status striving data

.739	.054	2.452	.058[1]

Suppose that you wanted to test the null hypothesis that these data arose from a population in which the underlying Pearson's product-moment correlation coefficient is 0, against the one-sided alternative that authoritarianism and social status striving are positively correlated. Using the techniques described in Chapter 1, you see that the asymptotic two-sided p value is 0.058. In contrast, the exact one-sided p value is 0.037. You can conclude that the asymptotic method does not perform well in this small data set.

Data With Ties. The diastolic blood pressure (mm Hg) was measured on 6 subjects in a treatment group and 7 subjects in a control group. The data are shown in Figure 2.17.

Figure 2.17 Diastolic blood pressure of treated and control groups



	pressure	group
1	94	Treated
2	108	Treated
3	110	Treated
4	90	Treated
5	108	Treated
6	105	Treated
7	80	Control
8	94	Control
9	94	Control
10	90	Control
11	90	Control
12	94	Control
13	94	Control

The results of the two-sample Kolmogorov-Smirnov test for these data are shown in Figure 2.18.

The asymptotic two-sided *p* value is 0.113. In contrast, the exact two-sided *p*

The two-sample Kolmogorov-Smirnov results for these data, without ties, are shown in Figure 2.20.

The asymptotic Kolmogorov-Smirnov two-sided *p* value remains unchanged at 0.113. This time, however, it is much closer to the exact two-sided *p* value, which is 0.091.

Large but Unbalanced Data Sets

Data from a prospective study of maternal drinking and congenital sex organ malformations (Graubard and Korn, 1987) are shown in Figure 2.21 in the form of a contingency table.

The linear-by-linear association test may be used to determine if there is a dose-response relationship between the average number of drinks consumed each day during pregnancy, and the presence of a congenital sex organ malfor

Sparse Data Sets

Data were gathered from 250 college and university administrators on various indicators

Figure 2.24 shows the asymptotic results of the Pearson chi-square test for these data.

Figure 2.24 Monte Carlo results for student/faculty ratio vs. competitiveness data

The asymptotic *p* value based on the Pearson chi-square test is 0.039, suggesting that there is an interaction between competitiveness and the student/faculty ratio. Notice,

3 One-Sample Goodness-of-Fit Inference

This chapter discusses tests used to determine how well a data set is fitted by a specified distribution. Such tests are known as goodness-of-fit tests. Exact Tests computes exact and asymptotic *p* values for the chi-square and Kolmogorov-Smirnov tests.

Available Tests

Table 3.1 shows the goodness-of-fit tests available in Exact Tests, the procedure from which each can be obtained, and a bibliographical reference for each.

Table 3.1 Available tests

Test	Procedure	References
Chi-square	Nonparametric Tests: Chi-square	Siegel and Castellan (1988)
Kolmogorov-Smirnov	Nonparametric Tests: 1 Sample K-S	Conover (1980)

Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit test is applicable either to categorical data or to continuous data that have been pre-grouped into a discrete number of categories. In tabular form, the data are organized as a $1 \quad c$ contingency table, where c is the number of categories. Cell i of this $1 \quad c$ table contains a frequency count, O_i, of the number of observations falling into category i. Along the bottom of the table is a $(1 \quad c)$ vector of cell probabilities

$$= (\ _1, \ _2, - \ _c) \qquad \text{Equation 3.1}$$

such that $_i$ is associated with column i. This representation is shown in Table 3.2

Table 3.2 Frequency counts for chi-square goodness-of-fit test

	Multinomial Categories				**Row Total**
	1	**2**	**...**		
Cell Counts	O_1	O_2	...	O_c	N
Cell Probabilities	1	2	...	c	1

The chi-square goodness-of-fit test is used to determine with judging if the data arose by taking N independent samples from a multinomial distribution consisting of c categories with cell probabilities given by . The null hypothesis

$$H_0:(O_1, O_2, \_\ O_c) \quad \text{Multinomial}(\ , N) \qquad \text{Equation 3.2}$$

can be tested versus the general alternative that H_0 is not true. The test statistic for the test is

$$X^2 = \sum_{i=1}^{c} (O_i - E_i)^2 / E_i \qquad \text{Equation 3.3}$$

where $E_i = N\ _i$ is the expected count in cell i. High values of X^2 indicate lack of fit and lead to rejection of H_0. If H_0 is true, asymptotically, as N , the random variable X^2 converges in distribution to a chi-square distribution with $(c-1)$ degrees of freedom. The asymptotic p value is, therefore, given by the right tail of this distribution. Thus, if x^2 is the observed value of the test statistic X^2, the asymptotic two-sided p value is given by

$$\tilde{p}_2 = \Pr(\ ^2_{c-1} \geq \tilde{x}^2) \qquad \text{Equation 3.4}$$

The asymptotic approximation may not be reliable when the E_i's are small. For example, Siegel and Castellan (1988) suggest that one can safely use the approximation only if at least 20% of the E_i's equal or exceed 5 and none of the E_i's are less than 1. In cases where the asymptotic approximation is suspect, the usual procedure has been to collapse categories to meet criteria such as those suggested by Siegel and Castellan. However, this introduces subjectivity into the analysis, since differing p values can be obtained by using different collapsing schemes. Exact Tests gives the exact p values without making any assumptions about the $_i$'s or N.

The exact *p* value is computed in Exact Tests by generating the true distribution of X^2 under H_0. Since there is no approximation, there is no need to collapse categories, and the natural categories for the data can be maintained. Thus, the exact two-sided *p* value is given by

$$p_2 = \Pr(\ ^2 \geq \tilde{x}^2) \qquad \text{Equation 3.5}$$

Sometimes a data set is too large for the exact *p* value to be computed, yet there might be reasons why the asymptotic *p* value is not sufficiently accurate. For these situations, Exact Tests provides a Monte Carlo estimate of the exact *p* value. This estimate is obtained by generating *M* multinomial vectors from the null distribution and counting how many of them result in a test statistic whose value equals or exceeds x^2, the test statistic actually observed. Suppose that this number is *m*. If so, a Monte Carlo estimate of p_2 is

$$\hat{p}_2 = m / M \qquad \text{Equation 3.6}$$

A 99% confidence interval for p_2 is then obtained by standard binomial theory as

$$CI = \hat{p}_2 \quad 2.576\sqrt{(\hat{p}_2)(1-\hat{p}_2)/M} \qquad \text{Equation 3.7}$$

A technical difficulty arises when either $\hat{p}_2 = 0$ or $\hat{p}_2 = 1$. Now the sample standard deviation is 0, but the data do not support a confidence interval of zero width. An alternative way to compute a confidence interval that does not depend on is based on inverting an exact binomial hypothesis test when an extreme outcome is encountered. If $\hat{p}$, an confidence interval for the exact *p* value is

Equation 3.8

Similarly, when , an confidence interval for the exact *p* value is

Equation 3.9

Exact Tests uses default values of and . While these defaults can be easily changed, they provide quick and accurate estimates of exact *p* values for a wide range of data sets.

Example: A Small Data Set

Table 3.3 shows the observed counts and the multinomial probabilities under the null hypothesis for a multinomial distribution with four categories.

Table 3.3 Frequency counts from a multinomial distribution with four categories

	Multinomial Categories				Row Total
	1	**2**	**3**	**4**	
Cell Counts	7	1	1	1	10
Cell Probabilities	0.3	0.3	0.3	.0.1	1

The results of the exact chi-square goodness-of-fit test are shown in Figure 3.1

Figure 3.1 Chi-square goodness-of-fit results

CATEGORY

	Observed N	Expected N	Residual
1	7	3.0	4.0
2	1	3.0	-2.0
3	1	3.0	-2.0
4	1	1.0	.0
Total	10		

Test Statistics

	Chi-Square[1]	df	Asymp. Sig.	Exact Sig.	Point Probability
CATEGORY	8.000	3	.046	.052	.020

1. 4 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.0.

The value of the chi-square goodness-of-fit statistic is 8.0. Referring this value to a chi-square distribution with 3 degrees of freedom yields an asymptotic p value

$$\tilde{p}_2 = (\Pr \chi^2_3 \geq 8.0) = 0.046$$

However, there are many cells with small counts in the observed 1 4 contingency table. Thus, the asymptotic approximation is not reliable. In fact, the exact p value is

$$p_2 = \Pr(\chi^2 \geq 8.0) = 0.0523$$

significant at the 5% level, 100,000 multinomial vectors can be sampled from the null distribution. The results are shown in Figure 3.3.

This time, the Monte Carlo estimate is 0.0508, almost indistinguishable from the exact result. Moreover, the exact *p*

Figure 3.4 Chi-square goodness-of-fit results for medium-sized data set

Multinomial Categories

	Observed N	Expected N	Residual
1	12	10.0	2.0
2	7	15.0	-8.0
3	31	25.0	6.0
Total	50		

Test Statistics

	Multinomial Categories
Chi-Square[1]	6.107
df	2
Asymp. Sig.	.047
Exact Sig.	.051
Point Probability	.002

1. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 10.0.

Notice that the asymptotic approximation gives a p value of 0.0472, while the exact p value is 0.0507. Thus, at the 5% significance level, the asymptotic value erroneously leads to rejection of the null hypothesis, despite the reasonably large sample size, the small number of categories, and the fact that $E_i \geq 10$ for $i = 1, 2, 3$.

One-Sample Kolmogorov Goodness-of-Fit Test

The one-sample Kolmogorov test is used to determine if it is reasonable to model a data set consisting of independent identically distributed (i.i.d.) observations from a completely specified distribution. Exact tests offers this test for the normal, uniform, and Poisson distributions.

Example: Testing for a Uniform Distribution

This example is taken from Conover (1980). A random sample size of 10 is drawn from

Runs Test

Consider a sequence of *N* binary outcomes, , where each is either a 0 or a 1. A run is defined as a succession of identical numbers that are followed and preceded by a different number, or no number at all. For example, the sequence

(1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1)

begins with a run of two 1's. A run of three 0's follows, and next a run of one 1. Then comes a run of four 0's, followed by a run of two 1's which in turn is followed by a run

Suppose that r is the observed value of the random variable R. The two-sided exact p value is defined as

$$p_2 = (\Pr|R - E(R)| \geq |r - E(R)|) \qquad \text{Equation 4.7}$$

where $E(R)$ is the expected value of R.

If a data set is too large for the computation shown in Equation 4.7 to be feasible, these p values can be estimated very accurately using Monte Carlo sampling.

For large data sets, asymptotic normality can be invoked. Let r denote the observed value of the random variable R, $h = 0.5$ if $r < (2mn/N) + 1$, and $h = -0.5$ if $r > (2mn/N) + 1$. Then the statistic

$$z \qquad \frac{r + h - (2mn/N) - 1}{[2mn(2mn - N)]} \qquad \text{Equation 4.8}$$

is normally distributed with a mean of 0 and a variance of 1.

The above exact, Monte Carlo, and asymptotic results apply only to binary data. However, you might want to test for the randomness of any general data series , where the 's are not binary. In that case, the approach suggested by Lehmann (1975) is to replace each with a corresponding binary transformation

Equation 4.9

where is the median of the observed data series. The median is calculated in the following way. Let be the observed data series sorted in ascending order. Then

Equation 4.10

Once this binary transformation has been made, the runs test can be applied to the binary data, as illustrated in the following data set. In addition to the median, the mean, mode, or any specified value can be selected as the cut-off for the runs test.

Example: Children's Aggression Scores

Figure 4.2 displays in the Data Editor the aggression scores for 24 children from a study of the dynamics of aggression in young children. These data appear in Siegel and Castellan (1988).

Figure 4.2 Aggression scores in order of occurrence

child	score
1	31
2	23
3	36
4	43
5	51
6	44
7	12
8	26
9	43
10	75
11	2
12	3
13	15
14	18
15	78
16	24
17	13
18	27
19	86
20	61
21	13
22	7
23	6
24	8

Figure 4.3 shows the results of the runs test for these data.

Figure 4.3 Runs test results for aggression scores data

	Test Value[1]	Cases < Test Value	Cases >= Test Value	Total Cases	Number of Runs	Z	Asymp. Sig. (2-tailed)	Exact Significance (2-tailed)	Point Probability
SCORE	25.00	12	12	24	10	-1.044	.297	.301	.081

1. Median

To obtain these results, Exact Tests uses the median of the 24 observed scores (25.0) as the cut-off for transforming the data into a binary sequence in accordance with Equation 4.8. This yields the binary sequence

(1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0).

Notice that this binary sequence of 12 1's and 12 0's contains 10 runs. Exact Tests determines that all permutations of the 12 1's and 12 0's would yield anywhere between a minimum of 2 runs and a maximum of 24 runs. The exact two-sided p value, or

	Test Value[1]	Cases < Test Value	Cases >= Test Value	Total Cases	Number of Runs	Z	Asymp. Sig. (2-tailed)	Exact Significance (2-tailed)	Point Probability
SCORE	1.00	4	6	10	3	-1.616	.106	.071	.038

Notice that the asymptotic two-sided *p* value is 0.106, while the exact two-sided *p* value is 0.071.

Two-Sample Inference: Paired Samples

The tests in this section are commonly applied to matched pairs of data, such as when

When to Use Each Test

The tests in this chapter have the common feature that they are applicable to data sets consisting of pairs of correlated data. The goal is to test if the first member of the pair has a different probability distribution from the second member. The choice of test is primarily determined by the type of data being tested: continuous, binary, or categorical.

Sign test. This test is used when observations in the form of paired responses arise from continuous distributions (possibly with ties), but the actual data are not available to us. Instead, all that is provided is the sign (positive or negative) of the difference in responses of the two members of each pair.

Wilcoxon signed-ranks test. This test is also used when observations in the form of paired responses arise from continuous distributions (possibly with ties). However, you now have the sign of the difference. You also have its rank in the full sample of response differences. If this additional information is available, the Wilcoxon signed-ranks test is more powerful than the sign test.

McNemar test.

Statistical Methods

For all the tests in this chapter, the data consist of correlated pairs of observations. For some tests, the observations are continuous (possibly with ties), while for others the observations are categorical. Nevertheless, in all cases, the goal is to test the null hypothesis that the two populations generating each pair of observations are identical. The basic permutation argument for testing this hypothesis is the same for all the tests. By this argument, if the null hypothesis were true, the first and second members of each pair of observations could just as well have arisen in the reverse order. Thus, each pair can be permuted in two ways, and if there are N pairs of observations, there are equally likely ways to permute the data. By actually carrying out these permutations, you can obtain the exact distribution of any test statistic defined on the data.

Sign Test and Wilcoxon Signed-Ranks Test

The data consist of N paired observations , where the X and Y random variables are correlated, usually through a matched-pairs design. Define the N differences

Omit from further consideration all pairs with a zero difference. Assume that for all . The following assumptions are made about the distribution of the random variables :

1. The distribution of each is symmetric.
2. The 's are mutually independent.
3. The 's have the same median.

Let the common median of the N 's be denoted by

whose observed value is

Equation 5.5

In other words, is the count of the number of positive differences among the N differences.

The permutational distributions of and under the null hypothesis are

For notational convenience, you can drop the subscript and let T denote either the statistic for the sign test or the statistic for the Wilcoxon signed-ranks test. The p value computations that follow are identical for both tests, with the understanding that T denotes when the Wilcoxon signed-ranks test is being computed and denotes when the sign test is being computed. In either case, you can now denote the standardized test statistic as

Equation 5.11

The two-sided asymptotic p value is defined, by the symmetry of the normal distribution, to be double the one-sided p value:

Equation 5.12

The exact one-sided p value is defined as

Equation 5.13

where t is the observed value of T

An unbiased Monte Carlo point estimate of the one-sided *p* value is

Equation 5.15

Next, if , so that you are estimating the left tail of exact distribution, the random variable is defined by

The Monte Carlo point estimate of the one-sided *p* value is once again given by Equation 5.15.

A 99% confidence interval for the exact one-sided *p* value is

Equation 5.16

The constant in the above equation, 2.576, is the upper 0.005 quantile of the standard normal distribution. It arises because Exact Tests chooses a 99% confidence interval for the *p* value as its default. However, you can easily choose any confidence level for the Monte Carlo estimate of the *p* value. Ordinarily, you would not want to lower the level of the Monte Carlo confidence interval to below the 99% default, since there should be a high assurance that the exact *p* value is contained in the confidence interval.

A technical difficulty arises when either or . Now the sample standard deviation is 0, but the data do not support a confidence interval of zero width. An alter-

natt92uTw[(d)Tw[(n) TD0.00 0 TD0.00c.2-2518.1(. No)-5i]vm11.5to27 3 -1.1627 -6 27 3 -14(t)-5

You can show that the variance of the two-sided Monte Carlo *p* value is four times as large as the variance of the corresponding one.86ued Monte Carlo

The Monte Carlo point estimate of the exact one-sided *p* value is 0.001, very close to the

Figure 5.4 Sign test results for AZT data

Frequencies

		N
Serum Antigen Level Post AZT - Serum Antigen Level (pg/ml) Pre-AZT	Negative Differences[1]	2
	Positive Differences[2]	14
	Ties[3]	4
	Total	20

1. Serum Antigen Level Post AZT < Serum Antigen Level (pg/ml) Pre-AZT
2. Serum Antigen Level Post AZT > Serum Antigen Level (pg/ml) Pre-AZT
3. Serum Antigen Level Post AZT = Serum Antigen Level (pg/ml) Pre-AZT

Statistics

		Exact Significance (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Pairs	Serum Antigen Level Post AZT - Serum Antigen Level (pg/ml) Pre-AZT	.004[2,3]	.002[2]	.002

The exact one-sided p value is 0.002. Notice that the exact one-sided p value for the sign test, while still extremely significant, is nevertheless larger than the corresponding exact one-sided p value for the Wilcoxon signed-ranks test. Since the sign test only takes into account the signs of the differences and not their ranks, it has less power than the Wilcoxon signed-ranks test. This accounts for its higher exact p value. The corresponding asymptotic inference fails to capture this distinction.

McNemar Test

The McNemar test (Siegel and Castellan, 1988; Agresti, 1990) is used to test the equality of binary response rates from two populations in which the data consist of paired, dependent responses, one from each population. It is typically used in a repeated measurements situation in which each subject's response is elicited twice, once before and once after a specified event (treatment) occurs. The test then determines if the initial response rate (before the event) equals the final response rate (after the event). Suppose two binomial responses are observed on each of N individuals. Let y_{11} be the count of the number of individuals whose first and second responses are both positive. Let y_{22} be the count of the number of individuals whose first and second responses are both negative. Let y_{12} be the count of the number of individuals whose first response is positive and whose second response is negative. Finally, let y_{21} be the count of the number of individuals whose first response is negative and whose second response is positive. Then the McNemar test is defined on a single 2 2 table of the form

$$y = \begin{matrix} y_{11} & y_{12} \\ y_{12} & y_{22} \end{matrix}$$

Let ($_{11}$, $_{12}$, $_{21}$, $_{22}$) denote the four cell probabilities for this table. The null hypothesis of interest is

$$H_0: \ _{12} = \ _{21}$$

The McNemar test depends only on the values of the off-diagonal elements of the 2 2 table. Its test statistic is

$$MC(\mathrm{y}) = y_{12} - y_{21} \qquad \text{Equation 5.20}$$

Now let y represent any generic 2 2 contingency table, and suppose that x is the 2 2 table actually observed. The exact permutation distribution of the test statistic (see Equation 5.20) is obtained by conditioning on the observed sum of off-diagonal terms, or **discordant pairs**,

$$N_d = y_{12} + y_{21}$$

The reference set is defined by

$$\Gamma = \{ y : y \text{ is } 2 \quad 2; y_{12} + y_{21} = N_d \} \qquad \text{Equation 5.21}$$

Under the null hypothesis, the conditional probability, , of observing any is binomial with parameters . Thus,

Equation 5.22

and the probability that the McNemar statistic equals or exceeds its observed value , is readily evaluated as

Equation 5.23

the sum being taken over all . The probability that the McNemar statistic is less than or equal to is similarly obtained. The exact one-sided *p* value is then defined as

Equation 5.24

You can show that the exact distribution of the test statistic is symmetric about 0. Therefore, the exact two-sided *p* value is defined as double the exact one-sided *p* value:

Equation 5.25

In large samples, the two-sided asymptotic p value is calculated by a approximation with a continuity correction, and 1 degree of freedom, as shown in Equation 5.26.

Equation 5.26

The definition of the one-sided *p* value for the exact case as the minimum of the left and right tails must be interpreted with caution. It should not be concluded automatically, based on a small one-sided *p* value, that the data have yielded a statistically significant outcome in the direction originally hypothesized. It is possible that the population difference occurs in the opposite direction from what was hypothesized before gathering the data. The direction of the difference can be determined from the sign of the test statistic, calculated as shown in Equation 5.27.

Equation 5.27

You should examine the one-sided *p* value as well as the sign of the test statistic before drawing conclusio7 Tc-0uul

Example: Voters' Preference

The following data are taken from Siegel and Castellan (1988). The crosstabulation shown in Figure 5.5 shows changes in preference for presidential candidates before and after a television debate.

Figure 5.5 Crosstabulation of preference for presidential candidates before and after TV debate

Preference Before TV Debate * Preference After TV Debate Crosstabulation

Count

		Preference After TV Debate	
		Carter	Reagan
Preference Before TV Debate	Carter	28	13
	Reagan	7	27

The results of the McNemar test for these data are shown in Figure 5.6.

Figure 5.6 McNemar test results

Test Statistics[1]

	N	Exact Significance (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Preference Before TV Debate & Preference After TV Debate	75	.263[2]	.132	.074

1. McNemar Test

The exact one-sided p value is 0.132. Notice that the value of the McNemar statistic, , has a positive sign. This indicates that of the 20 ( ) discordant pairs, more switched preferences from Carter to Reagan (13) than from Reagan to Carter (7). The point probability, 0.074, is the probability that .

The question of interest is whether there is agreement between the two pathologists. One way to answer this question is through the measures of association discussed in Part 4. Another way is to run the test of marginal homogeneity. The results of the exact marginal homogeneity test are shown in Equation 5.10.

Figure 5.10 Results of marginal homogeneity test

Marginal Homogeneity Test

	Distinct Values	Off-Diagonal Cases	Observed MH Statistic	Mean MH Statistic	Std. Deviation of MH Statistic	Std. MH Statistic	Asymp. Sig. (2-tailed)	Exact Significance (2-tailed)	Exact Sig. (1-tailed)	Point Probability
First Pathologist & Pathologist 2	5	43	114.000	118.500	3.905	-1.152	.249	.307	.154	.053

The exact two-sided *p* value is 0.307, indicating that the classifications by the two pathologists are not significantly different. Notice, however, that there is a fairly large difference between the exact and asymptotic *p* values because of the sparseness in the off-diagonal elements.

Two-Sample Inference: Independent Samples

This chapter discusses tests based on two independent samples of data drawn from two distinct populations. The objective is to test the null hypothesis that the two populations have the same response distributions against the alternative that the response distributions are different. The data could also arise in randomized clinical trials in which each subject is assigned randomly to one of two treatments. The goal is to test whether the treatments differ with respect to their response distributions. Here it is not necessary to mak67 -9mcE(ls in)-5( )-6(noycE(ls (makTc-0.0Twe7 -9mceeo-Sa)5.8(a)-0.5ooycE(o n1 g114tc

When to Use Each Test

The tests in this chapter deal with the comparison of samples drawn from the two distributions. The null hypothesis is that the two distributions are the same.

The choice of test depends on the type of alternative hypothesis you are interested in detecting.

Mann-Whitney test. The Mann-Whitney test, or Wilcoxon rank-sum test, is one of the most popular two-sample tests. It is generally used to detect "shift alternatives." That is, the two distributions have the same general shape, but one of them is shifted relative to the other by a constant amount under the alternative hypothesis. This test has an asymptotic relative efficiency of 95.5% relative to the Student's *t* test when the underlying populations are normal.

Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test is a distribution-free test for the equality of two distributions against the general alternative that they are different. Because this test attempts to detect any possible deviation from the null hypothesis, it will no

Table 6.2 One-way layout for two independent samples

Samples	
1	**2**
u_{11}	u_{12}
u_{21}	u_{22}
.	.
.	.
.	.
.	$u_{n_2 2}$
.	
.	
$u_{n_1 1}$	

denote the distribution from which the n_j observations displayed in column j of the one-way layout were drawn. The goal is to test the null hypothesis

$$H_0: F_1 = F_2 \qquad \text{Equation 6.2}$$

The observations in u are independent both within and across columns. In order to test H_0 by nonparametric methods, it is necessary to replace the original observations in the one-way layout with corresponding scores. These scores represent various ways of ranking the data in the pooled sample of size N. Different tests utilize different scores. Let w_{ij} be the score corresponding to u_{ij}. Then the one-way layout, in which the original data have been replaced by scores, is represented by Table 6.3.

Table 6.3 One-way layout with scores replacing original data

Samples	
1	**2**
w_{11}	w_{12}
w_{21}	w_{22}
.	.
.	.
.	.
.	$w_{n_2 2}$
.	
.	
$w_{n_1 1}$	

This table, denoted by w, displays the observed one-way layout of scores. Inference about H is based on comparing this observed one-way layout to others like it, in which the individual elements are the same but they occupy different rows and columns. In order to develop this idea more precisely, let the set W denote the collection of all pos-

sible two-column one-way layouts, with n_1 elements in column 1 and n_2 elements in column 2, whose members include *w* and all its permutations. The random variable $\tilde{w}$ is a **permutation** of *w* if it contains precisely the same scores as *w*, but these scores have been rearranged so that, for at least one $(i, j), (i', j')$ pair, the scores $w_{i,j}$ and $w_{i',j'}$ are interchanged.

Formally, let

$$W = \{\tilde{w}: \tilde{w} = w, \text{ or } \tilde{w} \text{ is a permutation of } w\} \qquad \text{Equation 6.3}$$

where $\tilde{w}$ is a random variable, and *w* is a specific value assumed by it.

To clarify these concepts, let us consider a simple numerical example. Let the original data come from two independent samples of size 5 and 3, respectively. These data are displayed as the one-way layout shown in Table 6.4.

Table 6.4 One-way layout of original data

Samples	
1	**2**
27	38
30	9
55	27
72	
18	

As you will see in "Mann-Whitney Test" on p. 83, in order to perform the Mann-Whitney test on these data, the original data must be replaced by their ranks. The one-way layout of observed scores, based on replacing the original data with their ranks, is displayed in Table 6.5.

Table 6.5 One-way layout with ranks replacing original data

Samples	
1	**2**
3.5	6
5	1
7	3.5
8	
2	

This one-way layout of ranks is denoted by *w*. It is the one actually observed. Notice that two observations were tied at 27 in *u*. Had they been separated by a small amount, they would have ranked 3 and 4. But since they are tied, the mid-rank $(3 + 4)/2 = 3.5$ is

used as the rank for each of them in *w*. The symbol *W* represents the set of all possible one-way layouts whose entries are the eight numbers in *w*, with five numbers in column 1 and three numbers in column 2. Thus, *w* is one member of *W*. (It is the one actually observed.) Another member is w', representing a different permutation of the numbers in *w*, as shown in Table 6.6.

The shift parameter is unknown. If it can be specified a priori that must be either

its variance is

$$\text{var}(T) \quad \frac{n_1 n_2}{12}\ n_1 \quad n_2 \quad 1 \quad \frac{\Sigma_{l=1}^{g} e_l(e_l^2 - 1)}{(n_1 + n_2)\ n_1 \quad n_2 - 1}$$

Equation 6.10

and its observed value is

Equation 6.11

The Wilcoxon rank-sum test statistic for the second column (or sample) is defined similarly.

In its Mann-Whitney form, this observed statistic is defined by subtracting off a constant:

Equation 6.12

The Wilcoxon rank-sum statistic corresponding to the column with the smaller Mann-Whitney statistic is displayed and used as the test statistic.

Exact P Values

The Wilcoxon rank-sum test statistic, *T,* is considered extreme if it is either very large or very small. Large values of *T* indicate a departure from the null hypothesis in the direction , while small values of *T* indicate a departure from the null hypothesis in the opposite direction, . Whenever the test statistic possesses a directional property of this type, it is possible to define both one- and two-sided *p* values. The exact one-sided *p* value is defined as

Equation 6.13

and the exact two-sided *p* value is defined as

Equation 6.14

Monte Carlo P Values

When exact *p* values are too difficult to compute, you can estimate them by Monte Carlo sampling. The following steps show how you can use Monte Carlo to estimate the exact *p* value given by Equation 6.14. The same procedure can be readily adapted to Equation 6.13.

1. Generate a new one-way layout of scores by permuting the original layout, *w*, in one of the $N!/(n_1!n_2!)$ equally likely ways.
2. Compute the value of the test statistic *T* for the permuted one-way layout.
3. Define the random variable

$$Z = \begin{cases} 1 \text{ if } |T - E(T)| \geq |t - E(T)| \\ 0 \text{ otherwise} \end{cases} \qquad \text{Equation 6.15}$$

Repeat the above steps a total of *M* times to generate the realizations $(z_1, z_2, \ldots z_M)$ for the random variable *Z*. Then an unbiased estimate of p_2 is

$$\hat{p}_2 = \frac{\sum_{l=1}^{M} z_l}{M} \qquad \text{Equation 6.16}$$

Next, let

$$\hat{} = \left[\frac{1}{M-1} \sum_{l=1}^{M} (z_l - \hat{p}_2)^2 \right]^{1/2} \qquad \text{Equation 6.17}$$

be the sample standard deviation of the z_l's. Then a 99% confidence interval for the exact *p* value is

$$CI = \hat{p}_2 \quad 2.576 \hat{} / \sqrt{M} \qquad \text{Equation 6.18}$$

A technical difficulty arises when either $\hat{p}_2 = 0$ or $\hat{p}_2 = 1$. Now the sample standard deviation is 0 but the data do not support a confidence interval of zero width. An alternative way to compute a confidence interval that does not depend on $\hat{}$ is based on inverting an exact binomial hypothesis test when an extreme outcome is encountered. It can be easily shown that if $\hat{p}_2 = 0$, an α% confidence interval for the exact *p* value is

$$CI = [0, 1 - (1 - \alpha/100)^{1/M}] \qquad \text{Equation 6.19}$$

Similarly, when $\hat{p}_2 = 1$, an α% confidence interval for the exact *p* value is

$$CI = [(1 - \alpha/100)^{1/M}, 1] \quad \text{Equation 6.20}$$

Exact Tests uses default values of $M = 10000$ and $\alpha = 99\%$. While these defaults can be easily changed, they provide quick and accurate estimates of exact *p* values for a wide range of data sets.

Asymptotic P Values

The one- and two-sided *p* values are obtained by computing the normal approximations to Equation 6.13 and Equation 6.14, respectively. Thus, the asymptotic one-sided *p* value is defined as

$$\tilde{p}_1 = \min\ \Phi((t - E(T))/\ _T), 1 - \Phi((t - E(T))/\ _T) \quad \text{Equation 6.21}$$

and the asymptotic two-sided *p* value is defined as

$$\tilde{p}_2 = 2\tilde{p}_1 \quad \text{Equation 6.22}$$

where $\Phi(z)$ is the tail area to the left of *z* from a standard normal distribution, and $_T$ is the standard deviation of *T*, obtained by taking the square root of 7.10.

Example: Blood Pressure Data

The diastolic blood pressure (mm Hg) was measured on 4 subjects in a treatment group and 11 subjects in a control group. Figure 6.1 shows the data displayed in the Data Editor. The data consist of two variables—*pressure* is the diastolic blood pressure of each subject, and *group* indicates whether the subject was in the experimentally *treated* group or the *control* group.

Figure 6.1 Diastolic blood pressure of treated and control groups

	pressure	group
1	94	Treated
2	108	Treated
3	110	Treated
4	90	Treated
5	80	Control
6	94	Control
7	85	Control
8	90	Control
9	90	Control

The Mann-Whitney test is computed for these data. The results are displayed in Figure 6.2.

Figure 6.2 Mann-Whitney results for diastolic blood pressure data

Ranks

			N	Mean Rank	Sum of Ranks
Diastolic Blood Pressure	Treatment Group	Treated	4	11.25	45.00
		Control	11	6.82	75.00
		Total	15		

Test Statistics[1]

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)	Exact Sig. [2*(1-tailed Sig.)]	Exact Significance (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Diastolic Blood Pressure	9.000	75.000	-1.720	.085	.104[2]	.099	.054	.019

1. Grouping Variable: Treatment Group

2. Not corrected for ties.

The Mann-Whitney statistic for the *treated* group, calculated by Equation 6.12, is 35.0 and for the *control* group is 9.0. Thus, the Wilcoxon rank-sum statistic for the control group is used. The observed Wilcoxon rank-sum statistic is 75. The Mann-Whitney U statistic is 9.0. The exact one-sided p value, 0.054, is not statistically significant at the 5% level. In this data set, the one-sided asymptotic p value, calculated as one-half of the two-sided p value, 0.085, is 0.0427. This value does not accurately represent the exact p value and would lead you to the erroneous conclusion that the treatment group is significantly different from

Figure 6.4 Monte Carlo results with 30,000 samples for diastolic blood pressure data

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)	Exact Sig. [2*(1-tailed Sig.)]	Sig.					
Diastolic Blood Pressure	9.000	75.000	-1.720	.085	.104[2]	.102[3]	.098	.107	.056[3]	.053	.059

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is applicable in more general settings than the Mann-Whitney test. Both are tests of the null hypothesis (see Equation 6.2). However, the Kolmogorov-Smirnov test is a universal test with good power against general alternatives in which and can differ in both shape and location. The Mann-Whitney test has good power against location shift alternatives of the form shown in Equation 6.7.

The Kolmogorov-Smirnov test is a two-sided test having good power against the alternative hypothesis

Equation 6.23

The Kolmogorov-Smirnov statistics used for testing the hypothesis in Equation 6.23 can now be defined. These statistics are all functions of the empirical cumulative density function (CDF) for and the empirical CDF for . "Statistical Methods" on p. 78 stated that the test statistics in this chapter are all functions of the one-way layout, *w*, displayed in Table 6.3, in which the original data have been replaced by appropriate scores. Indeed, this is true here as well, since you could use the original data as scores and construct an empirical CDF for each of the two samples of data. In that case, you would use as the one-way layout of scores. Alternatively, you could first convert the original data into ranks, just like those for the Mann-Whitney test, and then construct an empirical CDF for each of the two samples of ranked data. Hajek (1969) has demonstrated that in either case, the same inferences can be made. Thus, the Kolmogorov-Smirnov test is classified as a rank test. However, for the purpose of actually computing the empirical CDF's and deriving test statistics from them, it is often more convenient to work directly with raw data instead of first converting them into ranks (or mid-ranks, in the case of ties). Accordingly, let *u* be the actually observed one-

way layout of data, depicted in Table 6.2, and let *w*, the corresponding one-way layout of scores, also be *u*. Thus, the entries in Table 6.3 are the original 's. Now let ( ) denote the observations from the first sample sorted in ascending order, and let ( ) denote the observations from the second sample, sorted in ascending order. These sorted observations are often referred to as the order statistics of the sample. The empirical CDF for each distribution is computed from its order statistics. Before doing this, some additional notation is needed to account for the possibility of tied observations. Among the order statistics in the *j*th sample, , let there be distinct order statistics, with observations all tied for first place, observations all tied for second place, and so on until finally, observations are all tied for last place. Obviously, . Let ( ) represent the distinct order statistics of sample . You can now compute the empirical CDF's, for and for , as shown below. For , define

The test statistic for testing the null hypothesis (see Equation 6.2) against the two-sided alternative hypothesis (see Equation 6.23) is the Kolmogorov-Smirnov *Z* and is defined as

Equation 6.24

where *T* is defined as

Equation 6.25

and the observed value of *T* is denoted by *t*. The exact two-sided *p* value for testing Equation 6.2 against Equation 6.23 is

Equation 6.26

When the exact *p* value is too difficult to compute, you can resort to Monte Carlo sam-

3. Define the random variable

Equation 6.27

Repeat the above steps a total of M times to generate the realizations for the random variable Z. Then an unbiased estimate of is

Equation 6.28

Next, let

Equation 6.29

be the sample standard deviation of the 's. Then a 99% confidence interval for the exact p

The asymptotic two-sided *p* value, , is based on the following limit theorem:

Equation 6.33

Figure 6.6 Two-sample Kolmogorov-Smirnov results for orange juice and ascorbic acid data

Frequencies

			N
Score	Source of Vitamin C	Orange Juice	10
		Ascorbic Acid	10
		Total	20

Test Statistics[1]

		Score
Most Extreme Differences	Absolute	.600
	Positive	.000
	Negative	-.600
Kolmogorov-Smirnov Z		1.342
Asymp. Sig. (2-tailed)		.055
Exact Significance (2-tailed)		.045
Point Probability		.043

1. Grouping Variable: Source of Vitamin C

The exact two-sided p value is 0.045. This demonstrates that, despite the small sample size, there is a statistically significant difference between the two forms of vitamin C administration. The corresponding asymptotic p value equals 0.055, which is not statistically significant. It has been demonstrated in several independent studies (see, for example, Goodman, 1954) that the asymptotic result is conservative. This is borne out in the present example.

Wald-Wolfowitz Runs Test

The Wald-Wolfowitz runs test is a competitor to the Kolmogorov-Smirnov test for testing the null hypothesis

$$H_0: \; F_1(v) = F_2(v) \text{ for all } v \qquad \text{Equation 6.34}$$

against the alternative hypothesis

Equation 6.35

The test is completely general, in the sense that no distributional assumptions need to be made about and . Thus, it is referred to as an omnibus, or distribution-free, test.

Suppose the data consist of the one-way layout displayed as Table 6.2. The Wald-Wolfowitz test statistic is computed in the following steps:

1. Sort all observations in ascending order, and position them in a single row represented as .
2. Replace each observation in the above row with the sample identifier 1 if it came from the first sample and 2 if it came from the second sample.
3. A run is defined as a succession of identical numbers that are followed and preceded by a different number or no number at all. The test statistic, *T*, for the Wald-Wolfowitz test is the number of runs in the above row of 1's and 2's.

Under the null hypothesis, you expect the sorted list of observations to be well mixed with respect to the sample 1 and sample 2 identifiers. In that case, you will see a large

Figure 6.7 shows the data displayed in the Data Editor. *Salary* represents the starting salaries of nine court employees hired between 1975 and 1979, and *gender* indicates the gender of the employee.

Figure 6.7 Starting monthly salaries (in dollars) of nine court clerical workers

	salary	gender
1	525	Female
2	500	Female
3	550	Female
4	576	Female
5	458	Female
6	600	Female
7	700	Male
8	886	Male
9	600	Male

A quick visual inspection of these data reveals that in no case was a female paid a higher starting salary than a male hired for a comparable position. Consider these data to clarify how the Wald-Wolfowitz statistic is obtained.

The table below consists of two rows. The first row contains the nine observations sorted in ascending order. The second row contains the sample identifier for each observation: 1 if female and 2 if male.

458	500	525	550	576	600	600	700	886
1	1	1	1	1	1	2	2	2

By the above definition, there are only two runs in these data. Notice, however, that there is a tie in the data. One observation from the first sample and one from the second sample are both tied with a value of 600. Therefore, you could also represent the succession of observations and their sample identifiers as shown below.

458	500	525	550	576	600	600	700	886
1	1	1	1	1	2	1	2	2

Now there are four runs in the above succession of sample identifiers. First, there is a run of five 1's. Then a run of a single 2, followed by a run of a single 1. Finally, there is a run of two 2's.

The liberal value of the Wald-Wolfowitz test statistic is the one yielding the smallest number of runs after rearranging the ties in all possible ways. This is denoted by $t_{\min}$. The conservative value of the Wald-Wolfowitz test statistic is the one yielding the largest

7 K-Sample Inference: Related Samples

This chapter discusses tests based on *K* related samples, each of size *N*. It is a generalization of the paired-sample problem described in Chapter 5. The data consist of *N* independent vectors or *blocks* of observations in which there is dependence among the *K* components of each block. The dependence can arise in various ways. Here are a few examples:

There are *K* repeated measurements on each of N subjects, possibly at different time points, once after each of *K* treatments has been applied to the subject.

There are *K* subjects within each of *N* independent ma TcD-0.wK5.4(-0.5(on seof ea)a co, we a

[redacted] of data with *K* correlated observations within [redacted] ontinuous (possibly with ties). However, for the [redacted] ary. Various test statistics can be defined on this two-way [redacted] ese test statistics are defined on ranked data rather than on the [redacted] ingly, first replace the *K* observations, $(u_{i1}, u_{i2}, \_\ u_{iK})$ in block *i* with corresponding ranks, $(r_{i1}, r_{i2}, \_\ r_{iK})$. If there were no ties among these u_{ij}s , you would assign the first *K* integers $(1, 2, \_\ K)$, not necessarily in order, as the ranks of these *K* observations. If there are ties, you would assign the average rank or mid-rank to the tied observations. Specifically, suppose that the *K* observations of the first block take on e_1 distinct values, with d_{21} of the observations being equal to the smallest value, d_{22} to the next smallest, *d* to the third smallest, and so on. Similarly, the *K* observations in the second block take on distinct values, with of the observations being equal to the smallest value, to the next smallest, to the third smallest, and so on. Finally, the *K* observations in the *N*th block take on distinct values, with of the observations being equal to the smallest value, to the next smallest, to the third smallest, and so on. It is now possible to define the mid-ranks precisely. For , the distinct mid-ranks in the *i*th block, sorted in ascending order, are

Equation 7.1

You can now replace the original observations, , in the *i*th block with corresponding mid-ranks, , where each is the appropriate selection from the set of distinct mid-ranks . The modified two-way layout is shown in Table 7.3.

Table 7.3 Two-way layout for mid-ranks for K related samples

Block	**Treatments**		
Id	**1**	**2**	
1			
2			
.	.	.	.
.	.	.	.
.	.	.	.
.			
N			

As an example, suppose that $K = 5$, there are two blocks, and the two-way layout of the raw data (the ’s) is as shown in Table 7.4.

Table 7.4 Two-way layout w

For the first block, , with , , . Using Equation 7.1, you can obtain mid-ranks , , and . For the second block, , with , . Thus, you obtain mid-ranks and . You can now use these mid-ranks to replace the original values with corresponding values. The modified two-way layout, in which raw data have been replaced by mid-ranks, is displayed as Table 7.5.

All of the tests discussed in this chapter are based on test statistics that are functions of the two-way layout of mid-ranks displayed in Table 7.3. Before specifying these test statistics, define the rank-sum for any treatment j as

Equation 7.2

the average rank-sum for treatment j as

Equation 7.3

and the average rank-sum across all treatments as

Equation 7.4

The test statistics for Friedman's, Kendall's *W*, and Cochran's *Q* tests, respectively, are all functions of r_{ij}, $r_{.j}$, and $r_{..}$. The functional form for each test differs, and is defined later in this chapter in the specific section that deals with the test. However, regardless of its functional form, the exact probability distribution of each test statistic is obtained by the same permutation argument. This argument and the corresponding definitions of the one- and two-sided *p* values are given below.

Let *T* denote the test statistic for any of the tests in this chapter, and test the null hypothesis

H_0: There is no difference in the *K* treatments Equation 7.5

If H_0 is true, the *K* mid-ranks, $(r_{i1}, r_{i2}, \ldots r_{iK})$, belonging to block *i* could have been obtained in any order. That is, any treatment could have produced any mid-rank, and there are *K*! equally likely ways to assign the *K* mid-ranks to the *K* treatments. If you apply the same permutation argument to each of the *N* blocks, there are $(K!)^N$ equally likely ways to permute the observed mid-ranks such that the permutations are only carried out within each block but never across the different blocks. That is, there are $(K!)^N$ equally likely permutations of the original two-way layout of mid-ranks, where only intra-block permutations are allowed. Each of these permutations thus has a $(K!)^{-N}$ probability of being realized and leads to a specific value of the test statistic. The exact probability distribution of *T* can be evaluated by enumerating all of the permutations of the original two-way layout of mid-ranks. If *t* denotes the observed value of *T* in the original two-way layout, then

$$\Pr(T = t) = \sum_{T = t} (K!)^{-N} \qquad \text{Equation 7.6}$$

the sum being taken over all possible permutations of the original two-way layout of mid-ranks which are such that $T = t$. The probability distribution (see Equation 7.6) and its tail areas are obtained in Exact Tests by fast numerical algorithms. The exact two-sided *p* value is defined as

$$p_2 = \Pr(T \leq \hat{t}) = \sum_{T \leq \hat{t}} (K!)^{-N} \qquad \text{Equation 7.7}$$

When Equation 7.7 is too difficult to obtain by exact methods, it can be estimated by Monte Carlo sampling, as shown in the following steps:

1. Generate a new two-way layout of mid-ranks by permuting each of the *N* blocks of the original two-way layout of mid-ranks (see Table 7.3) in one of *K*! equally likely ways.

Friedman's test has good power against the alternative hypothesis

Equation 7.17

Notice that this alternative hypothesis is an omnibus one. It does not specify any ordering of the treatments in terms of increases in response levels. The alternative to the null hypothesis is simply that the treatments are different, not that one specific treatment is more effective than another.

Friedman's test uses the following test statistic, defined on the two-way layout of mid-ranks shown in Table 7.3.

Equation 7.18

The exact, Monte Carlo and asymptotic two-sided *p* values based on this statistic are obtained by Equation 7.7, Equation 7.9, and Equation 7.14, respectively.

Example: Effect of Hypnosis on Skin Potential

This example is based on an actual study (Lehmann, 1975). However, the original data have been altered to illustrate the importance of exact inference for data characterized by a small number of blocks but a large block size. In this study, hypnosis was used to elicit (in a random order) the emotions of fear, happiness, depression, calmness, and agitation from each of three subjects. Figure 7.1 shows these data displayed in the Data Editor. *Subject* identifies the subject, and *fear*, *happy*, *depress*, *calmness*, and *9and E4ple gis*

Do the five types of hypnotic treatments result in different skin measurements? The data seem to suggest that this is the case, but there were only three subjects in the sample. Friedman's test can be used to test this hypothesis accurately. The results are displayed in Figure 7.2.

Figure 7.2 Friedman's test results for hypnosis data

Ranks

	Mean Rank
FEAR	3.00
Happiness	5.00
Depression	1.50
Calmness	2.00
Agitation	3.50

Test Statistics[1]

N	3
Chi-Square	9.153
df	4
Asymp. Sig.	.057
Exact Sig.	.027
Point Probability	.003

1. Friedman Test

The exact two-sided *p* value is 0.027 and suggests that the five types of hypnosis are significantly different in their effects on skin potential. The asymptotic two-sided *p* value, 0.057, is double the exact two-sided *p* value and does not show statistical significance at the 5% level.

Because this data set is small, the exact computations can be executed quickly. For a larger data set, the Monte Carlo estimate of the exact *p* value is useful. Figure 7.3 displays the results of a Monte Carlo analysis on the same data set, based on generating 10,000 permutations of the original two-way layout.

Notice that the Monte Carlo point estimate of 0.027 is much closer to the true *p* value than the asymptotic *p* value. In addition, the Monte Carlo technique guarantees with 99% confidence that the true *p* value is contained within the range (0.023, 0.032). This confirms the results of the exact inference, that the differences in the five modes of hypnosis are statistically significant. The asymptotic analysis failed to demonstrate this result.

Kendall's W

Kendall's *W,*

Kendall's *W* bears a close relationship to Friedman's test; Kendall's *W* is in fact a

The point estimate of the coefficient of concordance is 0.656. The asymptotic *p* value of 0.055 suggests that you cannot reject the null hypothesis that the coefficient is 0. However, because of the small sample size (only 3 raters), this conclusion should be verified with an exact test, or you can rely on a Monte Carlo estimate of the exact *p* value, based

of Monte Carlo samples for both tests. If a different starting seed had been used, the two Monte Carlo estimates of the exact *p* value would have been slightly different.

Example: Relationship of Kendall's W to Spearman's R

In Chapter 14, a different measure of association known as Spearman's rank-order correlation coefficient is discussed. That measure is applicable only if there are judges, each ranking *K* applicants. Could this measure be extended if *N* exceeded 2? One approach might be to form distinct pairs of judges. Then each pair would yield a value for Spearman's rank-order correlation coefficient. Let denote the average of all these Spearman correl

Cochran's Q Test

Suppose that the u_{ij} values in the two-way layout shown in Table 7.2 were all binary, with a 1 denoting success and a 0 denoting failure. A popular mathematical model for generating such binary data in the context of the two-way layout is the logistic regression model

$$\log \frac{\quad_{ij}}{1 - \quad_{ij}} = \quad + \quad_i + \quad_j \qquad \text{Equation 7.21}$$

where, for all $i = 1, 2, \_ N$, and $j = 1, 2, \_ K$, $\quad_{ij} = \Pr(U_{ij} = 1)$, $\quad$ is the background log-odds of response, $\quad_i$ is the block effect, and $\quad_j$ is the treatment effect. All of these parameters are unknown, but for identifiability you can assume that

$$\sum_{i=1}^{N} \quad_i = \sum_{j=1}^{K} \quad_j = 0$$

Friedman's test applied to such data is known as Cochran's Q test. As before, the null hypothesis that there is no treatment effect can be formally stated as

$$H_0: (\quad_1 = \quad_2 = \_ = \quad_K) \qquad \text{Equation 7.22}$$

Cochran's Q test is used to test H_0 against unordered alternatives of the form

$$H_1: \quad_{j_1} \quad \quad_{j_2} \text{ for at least one } (j_1, j_2) \text{ pair} \qquad \text{Equation 7.23}$$

Like Friedman's test, Cochran's Q is an omnibus test. The alternative hypothesis is simply that the treatments are different, not that one specific treatment is more effective than another. You can use the same test statistic as for Friedman's test. Because of the binary observations, the test statistic reduces to

$$Q = \frac{K(K-1) \sum_{j=1}^{K} (B_j - \bar{B})^2}{K \sum_{l=1}^{N} L_i - \sum_{i=1}^{N} L_i^2} \qquad \text{Equation 7.24}$$

where B_j is the total number of successes in the jth treatment, L_i is the total number of successes in the ith block, and $\bar{B}$ denotes the average $(B_1 + B_2 + \_ + B_K)/K$. The asymptotic distribution of Q is chi-square with $(K-1)$ degrees of freedom. The exact

and Monte Carlo results are calculated using the same permutational arguments used for Friedman's test. The exact, Monte Carlo and asymptotic two-sided p values are thus obtained by Equation 7.7, Equation 7.9, and Equation 7.14, respectively.

Example: Crossover Clinical Trial of Analgesic Efficacy

This data set is taken from a three-treatment, three-period crossover clinical trial published by Snapinn and Small (1986). Twelve subjects each received, in random order, three treatments for pain relief: a placebo, an aspirin, and an experimental drug. The outcome of treatment j on subject i is denoted as either a success $u_{ij} = 1$ or a failure . Figure 7.7 shows the data displayed in the Data Editor.

This time, the exact *p* value, 0.059, is not significant at the 5% level, but the asymptotic approximation, 0.045, is. Although not strictly necessary for this small data set, you can also run the Monte Carlo test on the first 11 subjects. The results are shown in Figure 7.10.

Figure 7.10 Monte Carlo results for reduced analgesic efficacy data

Test Statistics

N	Cochran's Q	df	Asymp. Sig.	Monte Carlo Sig.		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
11	6.222[1]	2	.045	.056[2]	.050	.061

1. 0 is treated as a success.
2. Based on 10000 sampled tables with starting seed 2000000.

The Monte Carlo estimate of the exact *p* value was obtained by taking 10,000 random permutations of the observed two-way layout. As Figure 7.10 shows, the results matched those obtained from the exact test. The Monte Carlo sampling demonstrated that the exact *p* value lies in the interval (0.050, 0.061) with 99% confidence. This is compatible with the exact results, which also showed that the exact *p* value exceeds 0.05. The asymptotic result, on the other hand, erroneously claimed that the *p* value is less than 0.05 and is therefore statistically significant at the 5% level.

8 K-Sample Inference: Independent Samples

Table 8.1 Available tests

The Kruskal-Wallis and the Jonckheere-Terpstra

direction in which the *K* populations might be ordered under the alternative hypothesis. Such tests are said to be inherently two-sided.

Median test. The median test is useful when you have no idea whatsoever about the alternative hypothesis. It is an omnibus test for the equality of *K* distributions, where the alternative hypothesis is simply that the distributions are unequal, without any further specification as to whether they differ in shape, in location, or both. It uses only information about the magnitude of each of the observations relative to a single number, the median for the entire data set. Therefore, it is not as powerful as the other tests considered here, most of which use

Statistical Methods

The data for all the tests in this chapter consist of K independent samples each of size , where . These N observations can be represented in the form of the one-way layout shown in Table 8.2.

This table, denoted by u, shows the observed one-way layout of raw data. The observations in this one-way layout are independent both within and across columns. The data arise from continuous univariate distributions (possibly with ties). Let

Equation 8.1

denote the distribution from which the observations displayed in column j of the one-

Table 8.3 One-way layout with scores replacing original data

Samples			
1	**2**	…	
w_{11}	w_{12}	…	w_{1K}
w_{21}	w_{22}	…	2_{2K}
.	.	…	.
.	.		.
.	.		.
.	$w_{n_2 2}$	…	.
.			.
.			.
$w_{n_1 1}$			$w_{n_K K}$

This table, denoted by *w*, shows the observed one-way layout of scores. Inference about H_0 is based on comparing this observed one-way layout to others like it, in which the individual w_{ij} elements are the same but occupy different rows and columns. To develop this idea more precisely, let the set *W* denote the collection of all possible *K*-column one-way layouts, with n_j elements in column *j*, the members of which include *w* and all its permutations. The random variable $\tilde{w}$ is a permutation of *w* if it contains precisely the same scores as *w* but with the scores rearranged so that, for at least one $(i,j),(i',j')$ pair, the scores w_{ij} and $w_{w',j'}$ are interchanged. Formally, let

$$W = \{\tilde{w} : \tilde{w} = w, \text{ or } \tilde{w} \text{ is a permutation of } w\} \qquad \text{Equation 8.3}$$

In Equation 8.3, you could think of $\tilde{w}$ as a random variable, and *w* as a specific value assumed by it.

To clarify these concepts, consider a simple numerical example in which the original data come from three independent samples of size 5, 3, and 3, respectively. These data are displayed in a one-way layout, *u*, shown in Table 8.4.

Table 8.4 Example of a one-way layout of original data

Samples		
1	**2**	**3**
27	38	75
30	9	76
55	27	90
72		
18		

Distribution of T

In order to test the null hypothesis, , you need to derive the distribution of *T* under the assumption that is true. This distribution is obtained by the following permutational argument:

If is true, every member has the same probability of being observed.

Lehmann (1975) has shown that the above permutational argument is valid whether the data were gathered independently from *K* populations or were obtained by assigning*ua(men)-5 8Tr*

Exact P Values

For all tests against unordered alternatives, the more extreme values of T are those that are larger than the observed t. The exact two-sided p value is then defined as

Equation 8.7

Since there is no a priori natural ordering of the K treatments under the alternative hypothesis, large observed values of T are indicative of a departure from but not of the direction of the departure. Therefore, it is not possible to define a one-sided p value for tests against unordered alternatives.

For tests against ordered alternatives, such as the Jonckheere-Terpstra test, the test statistic T is considered extreme if it is either very large or very small. Large values of T indicate a departure from the null hypothesis in one direction, while small values of T indicate a departure from the null hypothesis in the opposite direction. Whenever the test statistic possesses a directional property of this type, it is possible to define both one- and two-sided p values. The exact one-sided p value is defined as

Equation 8.8

and the exact two-sided p value is defined as

Equation 8.9

where is the expected value of T.

Monte Carlo P Values

When exact p values are too difficult to compute, you can estimate them by Monte Carlo

p

Repeat the above steps a total of M times to generate the realizations $(z_1, z_2, \ldots z_M)$ for the random variable Z. Then an unbiased estimate of p_2 is

$$\hat{p}_2 = \frac{\sum_{l=1}^{M} z_l}{M}$$ Equation 8.11

Next, let

$$\hat{} = \left[\frac{1}{M-1} \sum_{l=1}^{M} (z_l - \hat{p}_2)^2 \right]^{1/2}$$ Equation 8.12

be the sample standard deviation of the z_l's . Then a 99% confidence interval for the exact p value is:

$$CI = \hat{p}_2 \quad 2.576 \hat{} / \sqrt{M}$$ Equation 8.13

A technical difficulty arises when either $\hat{p}_2 = 0$ or . Now the sample standard deviation is 0, but the data do not support a confidence interval of zero width. An alternative way to compute a confidence interval that does not depend on is based on inverting an exact binomial hypothesis test when an extreme outcome is encountered. It can be shown that if , an confidence interval for the exact p value is

Equation 8.14

Similarly when , an confidence interval for the exact p value is

Equation 8.15

Exact Tests uses default values of and . While these defaults can be easily changed, we have found that they provide quick and accurate estimates of exact p values for a wide range of data sets.

Asymptotic P Values

For tests against unordered alternatives the asymptotic two-sided p value is obtained by noting that the large-sample distribution of T is chi-square with degrees of freedom. The asymptotic p value is thus

Equation 8.16

As noted earlier, one–sided *p* values are not defined for tests against unordered alternatives.

For tests against ordered alternatives, in particular for the Jonckheere-Terpstra test, the asymptotic distribution of *T* is normal. The one- and two-sided *p* values are now defined by computing the normal approximations to Equation 8.8 and Equation 8.9, respectively. Thus, the asymptotic one-sided exact *p* value is defined as

Equation 8.17

and the asymptotic two-sided *p* value is defined as

Equation 8.18

The score, w_{ij}, corresponding to each u_{ij}, is defined as

Equation 8.21

Define

Equation 8.22

as the total number of observations in the *j*th sample that are at or below the median and

Equation 8.23

as the total number of observations in the pooled sample that are at or below the median.

The test statistic for the median test is defined on the contingency table displayed in Table 8.7. The entries in the first row are the counts of the number of subjects in each sample whose responses fall at or below the median, while the entries in the second row are the counts of the number of subjects whose responses fall above the median.

Table 8.7 Data grouped into a 2 x K contingency table for the median test

Group ID	**Samples**			**Row Total**
	1	**2**	**K**	
				m
Column Total				*N*

The probability of observing this contingency table under the null hypothesis, conditional on fixing the margins, is given by the hypergeometric function

Equation 8.24

For any , the test statistic for the median test is the usual Pearson chi-square statistic

Equation 8.25

Thus, if t is the value of T

Example: Hematologic Toxicity Data

The data on hematologic toxicity are shown in Figure 8.1. The data consist of two variables: *drug* is the chemotherapy regimen for each patient and *days* represents the number of days the patient's white blood count (WBC) was less than 500. The data consist of 28 cases.

Figure 8.1 Data on hematologic toxicity

	drug	days
1	1	0
2	1	1
3	1	8
4	1	10
5	2	0
6	2	0
7	2	3
8	2	3
9	2	8
10	3	5
11	3	6
12	3	7
13	3	14
14	3	14

15	4	1

The exact results of the median test for these data are shown in Figure 8.2, and the results of the Monte Carlo estimate of the exact test, using 10,000 Monte Carlo samples, are shown in Figure 8.3.

The median for the pooled sample is 7.0. This results in the value 4.317 for the test statistic, based on Equation 8.25. The exact p value is 0.429 and does not provide any evidence that the five

confidence interval for the exact *p* value, (0.419, 0.444) also supports the conclusion that there is no significant difference in the distribution of WBC across the five drugs.

The following discussion shows the relationship between the median test and the Pearson chi-square test. The median of these data is 7.0. The data can be divided into two groups, with one group containing those cases with $WBC \leq 7$ and the other group containing those cases with $WBC > 7$. The crosstabulation of these two groups, divided by the median, with the five drug regimens, is shown in Figure 8.4.

Figure 8.4 Hematologic toxicity data grouped into a 2 x K contingency table for the median test

Count

GROUP	WBC <= 7	2	4	3	6	1
	WBC > 7	2	1	2	3	4

The results of the Pearson chi-square test are shown in Figure 8.5. Notice that the results are the same as those obtained by running the median test on the original one-way layout of data.

Kruskal-Wallis Test

The Kruskal-Wallis test (Siegel and Castellan, 1988) is a very popular nonparametric test for comparing *K* independent samples. When , it specializes to the Mann-Whitney test. The Kruskal-Wallis test has good power against shift alternatives. Specifically, you assume, as in Hollander and Wolfe (1973), that the one-way layout, *u*, shown in Table 8.2, was generated by the model

Equation 8.29

for all and . In this model, is the overall mean, is the treatment effect, and the 's are identically distributed unobservable error terms from an unknown distribution with a mean of 0. All parameters are unknown, but for identifiability, you can assume that

Equation 8.30

The null hypothesis of no treatment effect can be formally stated as

Equation 8.31

The Kruskal-Wallis test has good power against the alternative hypothesis

Equation 8.32

Notice that this alternative hypothesis does not specify any ordering of the treatments in terms of increases in response levels. The alternative to the null hypothesis is simply that the treatments are different, not that one specific treatment elicits greater response than another. If there were a natural ordering of treatments under the alternative hypothesis—if, that is, you could state a priori that the 's are ordered under the alternative hypoth-

Jonckheere-Terpstra Test

being equal to the second smallest value, distinct observations being equal to the third smallest value, and so on, until, finally, distinct observations are equal to the largest value. The variance of the Jonckheere-Terpstra statistic is

Now, let be the observed value of *T*. The exact, Monte Carlo, and asymptotic *p* values based on the Jonckheere-Terpstra statistic can be obtained as discussed in "P Value Calculations" on p. 123. The exact one- and two-sided *p* values are computed as in Equation 8.8 and Equation 8.9, respectively. The Monte Carlo two-sided *p* value is computed as in Equation 8.11, with an obvious modification to reflect the fact that you want to estimate the probability inside the region instead of the region . The Monte Carlo one-sided *p* value can be similarly defined. The asymptotic distriotCha6e regio

Figure 8.8 Jonckheere-Terpstra test results for O-ring incidents data

The Jonckheere-Terpstra test statistic is displayed in its standardized form

Equation 8.40

whose observed value is

Equation 8.41

The output shows that , , and . Therefore, . The exact one-sided *p* value is

Equation 8.42

The exact two-sided *p* value is

Equation 8.43

These definitions are completely equivalent to those given by Equation 8.8 and Equation 8.9, respectively. Asymptotic and Monte Carlo one- and two-sided *p* values can be similarly defined in terms of the standardized test statistic. Note that is asymptotically normal with zero mean and unit variance.

The exact one-sided *p* value of 0.012 reveals that there is indeed a statistically signif-darand3(mperatuc-048 numbered t)-2.0.0329 Tw809value mean

The main objective is to test whether the observed contingency table is consistent with the null hypothesis of independence of row and column classifications. Exact Tests computes both exact and asymptotic *p* values for many different tests of this hypothesis against various alternative hypotheses. These tests are grouped in a logical manner and are presented in the next three chapters, which discuss unordered, singly ordered, and doubly ordered contingency tables, respectively. Despite these differences, there is a unified underlying framework for performing the hypothesis tests in all three situations. This unifying framework is discussed below in terms of *p* value computations.

The *p* value of the observed contingency table is used to test the null hypothesis of no row-by-column interaction. Exact Tests provides three categories of *p* values for each test. The "gold standard" is the exact *p* value. When it can be computed, the exact *p* p

Defining the Reference Set

Throughout this chapter, *x* will be used to denote the contingency table actually observed, and *y* will denote any generic contingency table belonging to some well-defined reference set of contingency tables that could have been observed. The exact probability of observing any generic table *y* depends on the sampling scheme used to generate it. When both the row and column classifications are categorical, Agresti (1990) lists three sampling schemes that could give rise to *y*—full multinomial sampling,

Notwithstanding the availability of the network algorithms, a data set is sometimes too large for the exact *p* value to be feasible to compute. But it might be too sparse for the asymptotic *p* value to be reliable. For this situation, Exact Tests also provides a Monte Carlo option, where only a small proportion of the $r \quad c$ tables in Γ are sampled, and an unbiased estimate of the exact *p* value is obtained.

Monte Carlo Two-Sided P Values

The **Monte Carlo two-sided value** is a very close approximation to the exact two-sided *p* value, but it is much easier to compute. The examples in Chapter 10, Chapter 11, and Chapter 12 will show that, for all practical purposes, the Monte Carlo results can be used in place of the exact results whenever the latter are too difficult to compute. The Monte Carlo approach is a steady, reliable procedure that, unlike the exact approach, always takes up a predictable amount of computing time. While it does not produce the exact *p* value, it does produce a fairly tight confidence interval within which the exact *p* value is contained, with a high degree of confidence (usually 99%).

In the Monte Carlo method, a total of *M* tables is sampled from Γ, each table being sampled in proportion to its hypergeometric probability (see Equation 9.2). (Sampling tables in proportion to their probabilities is known as **crude Monte Carlo sampling**.)

For each table $y_j \quad \Gamma$ that is sampled, define the binary outcome $z_j = 1$ if $y_j \quad \Gamma^*$; 0 otherwise. The arithmetic average of all *M* of these z_j's is taken as the Monte Carlo point estimate of the exact two-sided *p* value:

$$\hat{p}_2 = \frac{1}{M} \sum_{j=1}^{M} z_j \qquad \text{Equation 9.6}$$

It is easy to show that $\hat{p}_2$ is an unbiased estimate of the exact two-sided *p* value. Next,

$$\hat{} = \left[\frac{1}{M-1} \sum_{j=1}^{M} (z_j - \hat{p}_2)^2 \right]^{1/2} \qquad \text{Equation 9.7}$$

is the sample standard deviation of the z_j's. Then a 99% confidence interval for the exact *p* value is

$$CI = \hat{p}_2 \quad 2.576 \hat{} / (\sqrt{M}) \qquad \text{Equation 9.8}$$

A technical difficulty arises when either $\hat{p}_2 = 0$ or $\hat{p}_2 = 1$. The sample standard deviation is now zero, but the data do not support a confidence interval of zero width. An alternative way to compute a confidence interval that does not depend on is based on inverting an exact binomial hypothesis test when an extreme outcome is encountered. It can be easily shown that if $\hat{p}_2 = 0$, an α% confidence interval for the exact *p* value is

$$CI = [0, 1 - (1 - \alpha/100)^{1/M}] \qquad \text{Equation 9.9}$$

Similarly, when $\hat{p}_2 = 1$, an α% confidence interval for the exact *p* value is

$$CI = [(1 - \alpha/100)^{1/M}, 1] \qquad \text{Equation 9.10}$$

Asymptotic Two-Sided P Values

For all the tests in this chapter, the test statistic $D(y)$ has an asymptotic chi-square distribution. The **asymptotic two-sided value** is obtained as

$$\tilde{p}_2 = \Pr(\chi^2 \geq D(x) | df) \qquad \text{Equation 9.11}$$

where χ^2 is a random variable with a chi-square distribution and *df* are the appropriate degrees of freedom. For tests on unordered $r \times c$ contingency tables, the degrees of freedom are $(r-1)(c-1)$; for tests on singly ordered $r \times c$ contingency tables, the degrees of freedom are $(r-1)$; and tests on doubly ordered contingency tables have one degree of freedom. Since the square root of a chi-square variate with one degree of freedom has a standard normal distribution, you can also work with normally distributed test statistics for the doubly ordered $r \times c$ contingency tables.

Unordered R x C Contingency Tables

The tests in this chapter are applicable to contingency tables whose rows and columns cannot be ordered in a natural way. In the absence of such an ordering, it is not possible to specify any particular direction for the alternative to the null hypothesis that the row and column classifications are independent. The tests considered here are appropriate

three tests are asymptotically equivalent. The research in this area is scant and has focused primarily on the question of which of the three asymptotic tests best matches its exact counterpart. (See, for example, Roscoe and Byars, 1971; Chapman, 1976; Agresti and Yang, 1987; Read and Cressie, 1988.) It is very likely that the Pearson chi-square asymptotic test converges to its exact counterpart the fastest. You can use the Exact Tests option to investigate this question and also to determine empirically which of the three exact tests has the most power against specific alternative hypotheses.

Statistical Methods

For the $r \times c$ contingency table shown in Table 9.1, π_{ij} denotes the probability that an observation will be classified as belonging to row i and column j. Define the marginal probabilities:

$$\pi_{i+} = \sum_{j=1}^{c} \pi_{ij}, \text{ for } i = 1, 2, \ldots, r$$

$$\pi_{+j} = \sum_{i=1}^{r} \pi_{ij}, \text{ for } j = 1, 2, \ldots, c$$

The Pearson chi-square test, the likelihood-ratio test, and Fisher's exact test are all appropriate for testing the null hypothesis

$$H_0: \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all } (i,j)\text{pairs} \qquad \text{Equation 10.1}$$

against the general (omnibus) alternative that Equation 10.1 does not hold. An alternative hypothesis of this form is of interest when there is no natural ordering of the rows and columns of the contingency table. Thus, these three tests are usually applied to unordered $r \times c$ contingency tables. Note that all three tests are inherently two-sided in the following sense. A large positive value of the test statistic is evidence that there is at least one (i,j) pair for which Equation 10.1 fails to hold, without specifying which pair.

If the sampling process generating the data is product multinomial, one set of marginal probabilities (the π_{i+}'s, say) will equal unity. Then H_0 reduces to the statement that the c multinomial probabilities are the same for all rows. In other words, the null hypothesis is equivalent to

$$H_0: \pi_{1j} = \pi_{2j} = \ldots = \pi_{rj} = \pi_{+j} \text{ for all } j = 1,2,\ldots c \qquad \text{Equation 10.2}$$

In practice, product multinomial sampling arises when *r* populations are compared and the observations from each population fall into *c* distinct categories. The null hypothesis is that the multinomial probability of falling in the *j*th category, , is the same for each population. The Pearson, likelihood-ratio, and Fisher's tests are most suitable when the *c* categories have no natural ordering (for example, geographic regions of the country). However, more powerf

The question of interest is whether the distribution of the site of the oral lesion is significantly different in the three geographic regions. The row and column classifications for this table are clearly unordered, making it an appropriate data set for either the Pearson, likelihood-ratio or Fisher's tests. The contingency table is so sparse that the usual chi-square asymptotic distribution with 16 degrees of freedom is not likely to yield accurate

exact *p* value is 0.027, showing that there is a significant interaction between the site of the lesion and the geographic region, but the asymptotic *p* value failed to demonstrate this. In this example, the asymptotic *p* value was more conservative than the exact *p* value.

Sometimes the data set is too large for an exact analysis, and the Monte Carlo method must be used instead. Figure 10.3 shows an unbiased estimate of the exact *p* value for the Pearson chi-square test based on a crude Monte Carlo sample of 10,000 tables from the reference set.

Figure 10.3 Monte Carlo results for oral lesions data

Chi-Square Tests

		Values					
					Monte Carlo Significance (2-tailed)		
						99% Confidence Interval	
		Value	df	Asymp. Sig. (2-tailed)	Sig.	Lower Bound	Upper Bound
Statistics	Pearson Chi-Square	22.099[1]	16	.140	.026[2]	.022	.030

1. 25 cells (92.6%) have expected count less than 5. The minimum expected count is .26.
2. Based on 10000 and seed 2000000 ...

The Monte Carlo method produces a 99% confidence interval for the exact *p* value. Thus, although the point estimate might change slightly if you resample with a different starting seed or a different random number generator, you can be 99% confident that the exact *p* value is contained in the interval 0.022 to 0.030. Moreover, you could always sample more tables from the reference set if you wanted to further narrow the width of this interval. Based on this analysis, it is evident that the Monte Carlo approach leads to the same conclusion as the exact approach, demonstrating that there is indeed a significant row-by-column interaction in this contingency table. The asymptotic inference failed to demonstrate any row-by-column interaction.

Likelihood-Ratio Test

The likelihood-ratio test is an alternative to the Pearson chi-square test for testing independence of row and column classifications in an unordered $r \quad c$ contingency table. For any observed $r \quad c$ contingency table, the test statistic, $D(x)$, is denoted as $LI(x)$ and is computed by the formula

$$LI(x) \quad 2 \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \qquad \text{Equation 10.4}$$

For the oral lesions data displayed in Figure 10.1, . The test statistic and its corresponding asymptotic and exact *p* values are shown in Figure 10.4.

The output shows that the observed value of the test statistic is . This statistic has an asymptotic chi-square distribution with 16 degrees of freedom. The asymptotic *p* value is computed as the area under the chi-square density function to the right of . The *p*

The Monte Carlo point estimate is 0.035, which is acceptably close to the exact *p* value of 0.036. More important, the Monte Carlo method also produces a confidence interval for the exact *p* value. Thus, although this point estimate might change slightly if you resample with a different starting seed or a different random number generator, you can be 99% confident that the exact *p* value is contained in the interval 0.030 to 0.039. Moreover, you could always sample more tables from the reference set if you wanted to further narrow the width of this interval. Based on this analysis, it is evident that the Monte Carlo approach leads to the same conclusion as the exact approach, demonstrating that there is indeed a significant row-by-column interaction in this contingency table. The asymptotic inference failed to demonstrate any row-by-column interaction.

Fisher's Exact Test

Fisher's exact test is traditionally associated with the single 2 2 contingency table. Its extension to unordered r c tables was first proposed by Freeman and Halton (1951). Thus, it is also known as the Freeman-Halton test. It is an alternative to the Pearson chi-square and likelihood-ratio tests for testing independence of row and column classifications in an unordered r c contingency table. Fisher's exact test is available for tables larger than 2 2 through the Exact Tests option. Asymptotic results are provided only for 2 2 tables, while exact and Monte Carlo results are available for larger tables. For any observed r c contingency table, the test statistic, $D(x)$, is denoted as $Fl(x)$ and is computed by the formula

$$FI(x) = -2\log(P(x)) \qquad \text{Equation 10.5}$$

where

$$= (2)^{(r-1)(c-1)/2} N^{-(rc-1)/2} \prod_{i=1}^{r} (m_i)^{(c-1)/2} \prod_{j=1}^{c} (n_j)^{(r-1)/2} \qquad \text{Equation 10.6}$$

For the oral lesions data displayed in Figure 10.1, $FI(x) = 19.72$. The exact *p* values are shown in Figure 10.6.

Figure 10.6 Fisher's exact test for oral lesions data

Chi-Square Tests

	Value	Exact Sig. (2-tailed)
Fisher's Exact Test	19.721	.010

The exact p value is defined by Equation 9.4 as the permutational probability

Singly Ordered R x C Contingency Tables

The test in this chapter is applicable to contingency tables in which the rows are unordered but the columns are ordered. This is a common setting, for example, when comparing *r* different drug treatments, each generating an ordered categorical response. It is assumed a priori that the treatments cannot be ordered according to their rate of effectiveness. If they can be ordered according to their rate of effectiveness—for example, if the treatments represent increasing doses of some drug—the tests in the next chapter are more applicable.

Available Test

Exact Tests offers the Kruskal-Wallis test for analyzing contingency tables in which the rows (

The null hypothesis is

$$H_0: \Pi_1 = \Pi_2 = \_ = \Pi_r \qquad \text{Equation 11.1}$$

The alternative hypothesis is that at least one set of multinomial probabilities is stochastically larger than at least one other set of multinomial probabilities. Specifically, for $i = 1, 2, \_, r$, let

$$_{ij} = \sum_{l=1}^{j} {}_{il}$$

The Kruskal-Wallis test is especially suited to detecting departures from the null hypothesis of the form

$$H_1: \text{for at least one } {}_{11} \qquad \text{Equation 11.2}$$

with strict inequality for at least one *j*. In other words, you want to reject when at least one of the populations is more responsive than the others.

Tumor Regression Rates Data

The tumor regression rates of five chemotherapy regimens, Cytoxan (CTX) alone, Cyclohexyl-chloroethyl nitrosurea (CCNU) alone, Methotrexate (MTX) alone, CTX+MTX, and CTX+CCNU+MTX were compared in a small clinical trial. Tumor regression was measured on a three-point scale: no response, partial response, or complete response. The crosstabulation of the results is shown in Figure 11.1.

Although Figure 11.1 shows the data in crosstabulated format to illustrate the concept of applying the Kruskal-Wallis test to singly ordered tables, this test is obtained from the Nonparametric Tests procedure, and your data must be structured appropriately for Nonparametric Tests. Figure 11.2 shows these data displayed in the Data Editor. The data consist of two variables. *Chemo* is a grouping variable that indicates the chemotherapy regimen, and *regressn* is an ordered categorical variable with three values, where 1=*No*

Figure 11.3 Results of Kruskal-Wallis test for tumor regression data

Ranks

			N	Mean Rank
Tumor Regression	Chemotherapy Regimen	CTMX	2	5.00
		CCNU	2	8.25
		MTX	3	5.00
		CTX+CCNU	4	8.25
		CTX+CCNU+MTX	6	13.08
		Total	17	

Test Statistics[1, 2]

	Chi-Square	df	Asymp. Sig.	Exact Sig.	Point Probability
Tumor Regression	8.682	4	.070	.039	.001

1. Kruskal Wallis Test
2. Grouping Variable: Chemotherapy Regimen

The observed value of the test statistic t, calculated by Equation 8.34, is 8.682. The asymptotic two-sided p value is based on the chi-square distribution with four degrees of freedom. The asymptotic p value is obtained as the area under the chi-square density function to the right of 8.682. This p value is 0.070. However, this p value is not reliable because of the sparseness of the observed contingency table.

The exact p value is defined by Equation 8.7 as the permutational probability $\Pr(T \geq 8.682 \mid y \in \Gamma)$. The exact p value is 0.039, which implies that there is a statistically significant difference between the five modes of chemotherapy. The asymptotic inference failed to demonstrate this. Below the exact p value is the point probability $\Pr(T = 8.682)$. This probability, 0.001, is a natural measure of the discreteness of the test statistic. Some statisticians recommend subtracting half of its value from the exact p value, in order to yield a less conservative mid-p value. (For more information on the role of the mid-p method in exact inference, see Lancaster, 1961; Pratt and Gibbons, 1981; and Miettinen, 1985.)

Sometimes the data set is too large for an exact analysis, and the Monte Carlo method must be used instead. Figure 11.4 shows an unbiased estimate of the exact p value for the Kruskal-Wallis test based on a crude Monte Carlo sample of 10,000 tables from the reference set.

The Monte Carlo point estimate is 0.043, which is practically the same as the exact *p* value of 0.039. Moreover, the Monte Carlo method also produces a confidence interval for the exact *p*

12 Doubly Ordered R x C Contingency Tables

The tests in this chapter are applicable to $r \quad c$ contingency tables in which both the rows and columns are ordered. A typical example would be an $r \quad c$ table obtained from a dose-response study. Here the rows (r) represent progressively increasing doses of some drug, and the columns (c) represent progressively worsening levels of drug toxicity. The goal is to test the null hypothesis that the response rates are the same at all dose levels. The tests in this chapter exploit the double ordering so as to have good power against alternative hypotheses in which an increase in the dose level leads to an increase in the toxicity level.

Available Tests

Exact Tests offers two tests for doubly ordered $r \quad c$ contingency tables: the Jonckheere-Terpstra test and the linear-by-linear association test. Asymptotically, both test statistics converge to the standard normal distribution or, equivalently, the squares of these statistics converge to the chi-square distribution with one degree of freedom. Both the exact and asymptotic p values are available from Exact Tests. The asymptotic p value is provided by default, while the exact p value must be specifically requested. If a data set is too large for the exact p value to be computed, Exact Tests offers a special option whereby the exact p value is estimated up to Monte Carlo accuracy. Although the logic of the Jonckheere-Terpstra test can be applied to doubly ordered contingency tables, this test is performed through the Nonparametric Tests: Tests for Several Independent Samples procedure. Table 12.1 shows the two available tests, the procedure from which each can be obtained, and a bibliographical reference to each test.

Table 12.1 Available tests

Test	Procedure	Reference
Jonckheere-Terpstra test	Nonparametric Tests: K Independent Samples	Lehmann (1973)
Linear-by-linear association test	Crosstabs	Agresti (1990)

In this chapter, the null and alternative hypotheses for these tests are specified, appropriate test statistics are defined, and each test is illustrated with a data set.

When to Use Each Test

The Jonckheere-Terpstra and linear-by-linear association tests, while not asymptotically equivalent, are competitors for testing row and column interaction in a doubly ordered table. There has been no formal statistical research on which test has greater power. Historically, the Jonckheere-Terpstra test was developed for testing continuous data in a nonparametric setting, while the linear-by-linear association test was used for testing categorical data in a loglinear models setting. However, either test is applicable for computing *p* values in contingency tables as long as both the rows and columns have a natural ordering. In this chapter, the Jonckheere-Terpstra test is applied to ordinal categorical data. See Chapter 8 for a discussion of using this test for continuous data. The linear-by-linear association test has some additional flexibility in weighting the ordering and in weighting the relative importance of successive rows or columns of the contingency table through a suitable choice of row and column scores. This flexibility is illustrated in the treatment of the numeri

for $i = 1, 2, ..., r$. Since the rows are ordered, it is possible to define one-sided alternative hypotheses of the form

$$H_1: \; _{1j} \leq \; _{2j} \leq \cdots \leq \; _{rj} \qquad \text{Equation 12.3}$$

or

$$H'_1: \; _{1j} \geq \; _{2j} \geq \cdots \geq \; _{rj} \qquad \text{Equation 12.4}$$

for $j = 1, 2, ..., c$, with strict inequality of at least one j. Both the Jonckheere-Terpstra and the linear-by-linear association tests are particularly appropriate for detecting departures from the null hypothesis of the form H_1 or H'_1, or for detecting the two-sided alternative hypothesis that either H_1 or H'_1 is true. Hypothesis H_1 implies that as you move from row i to row $(i + 1)$, the probability of the response falling in category $(j + 1)$ rather than in category j increases. Hypothesis H'_1 states the opposite, that as you move down a row, the probability of falling into the next higher category decreases. The test statistics for the Jonckheere-Terpstra and the linear-by-linear association tests are so defined that large positive values reject H_0 in favor of H_1, while large negative values reject in favor of .

Dose-Response Data

Patients were treated with a drug at four dose levels (100mg, 200mg, 300mg, 400mg) and then monitored for toxicity. The data are tabulated in Figure 12.1.

Notice that there is a natural ordering across both the rows and the columns of the above contingency table. There is also the suggestion that progressively increasing drug doses lead to increases in drug toxicity.

Jonckheere-Terpstra Test

Figure 12.1 shows the data in crosstabulated format to illustrate the concept of applying the Jonckheere-Terpstra test to doubly ordered tables, however this test is obtained from the Nonparametric Tests procedure, and your data must be structured appropriately for Nonparametric Tests. Figure 12.2 shows a portion of these data displayed in the Data Editor. The data consist of two variables. *Dose*

In the present example, the smaller permutational probability is the one that evaluates the right tail. It is displayed on the screen as $\Pr(T^* \geq 1.65) = 0.049$. The exact one-sided p value is the point probability $\Pr(T^* = 1.65)$. This probability, 0.000, is a natural measure of the discreteness of the test statistic. Some statisticians advocate subtracting half its value from the exact p value, thereby yielding a less conservative mid-p value. (See Lancaster, 1961; Pratt and Gibbons, 1981; and Miettinen, 1985 for more information on the role of the mid-p value in exact inference.) Equation 12.8 defines the exact two-sided p value

$$p_2 = \Pr(|T^*| \geq 1.648) = 0.100 \qquad \text{Equation 12.8}$$

Notice that this definition will produce the same answer as Equation 9.4, with $D(y) = (T^*(y))^2$ for all $y \in \Gamma$.

Sometimes the data set is too large for an exact analysis, and the Monte Carlo method must be used instead. Figure 12.4 displays an unbiased estimate of the exact one- and two-sided p value for the Jonckheere-Terpstra test based on a crude Monte Carlo sample of 10,000 tables from the reference set.

Figure 12.4 Monte Carlo results for Jonckheere-Terpstra test for dose-response data

4	227	9127.000	8827.500	181.760	1.648	.099	.101[2]

The Monte Carlo point estimate of the exact one-sided p value is 0.051, which is very close to the exact one-sided p value of 0.049. Moreover, the Monte Carlo method also produces a confidence interval for the exact p value. Thus, although this point estimate might change slightly if you resample with a different starting seed or a different random number generator, you can be 99% confident that the exact p value is contained in the interval 0.045 to 0.057. The Monte Carlo point estimate of the exact two-sided p value is 0.101, and the corresponding 99% confidence interval is 0.093 to 0.109. More tables could be sampled from the reference set to further narrow the widths of these intervals.

Linear-by-Linear Association Test

The linear-by-linear association test orders the tables in Γ according to the linear rank statistic. Thus, if the observed table is *x*, the unnormalized test statistic is

Equation 12.9

The upper portion of the output displays the asymptotic two-sided *p* value. The *p* values are evaluated as tail areas under a chi-square distribution. The standardized value for the linear-by-linear association test is . This value is normally distributed with

The Monte Carlo point estimate of the exact one-sided *p* value is 0.046, which is very close to the exact one-sided *p* value of 0.044. Moreover, the Monte Carlo method also produces a confidence interval for the exact *p* value. Thus, although this point estimate might change slightly if you resample with a different starting seed or a different random number generator, you can be 99% confident that the exact *p*

Figure 12.8 shows the results of the linear-by-linear association test on these scores.

Figure 12.8 Results of linear-by-linear association test on adjusted data

	Value	2-tai9101wr842383				
Linear-by-Linear Association	3.008[2]	1	.083	.078	.050	.005

Observe now that the one-sided asymptotic *p* value is 0.042, , which is statistically significant, but that the one-sided exact *p* value (0.050) is not statistically significant at the 5% level. Inference based on asymptotic theory, with a rigid 5% criterion for claiming statistical significance, would therefore lead to an incorrect conclusion.

13 Measures of Association

This chapter introduces some definitions and notation needed to estimate, test, and interpret the various measures of association computed by Exact Tests. The methods discussed here provide the necessary background for the statistical procedures described in Chapter 14, Chapter 15, and Chapter 16.

Technically, there is a distinction between an actual measure of association, regarded as a population parameter, and its estimate from a finite sample. For example, the correlation coefficient is a population parameter in a bivariate normal distribution, whereas Pearson's product moment coefficient *R* is an estimate of , based on a finite sample from this distribution. However, in this chapter, the term "measure of association" will be used to refer to either a population parameter or an estimate from a finite sample, and it will be clear from the context which is intended. In particular, the formulas for the various measures of association discussed in this chapter refer to sample estimates and their associated standard errors, not to underlying population parameters. Formulas are not provided for the actual population parameters. For each measure of association, the following statistics are provided:

A point estimate for the measure of association (most often this will be the maximum-likelihood estimate [MLE]).

Its asymptotic standard error, evaluated at the maximum-likelihood estimate (ASE1).

Asymptotic two-sided *p* values for testing the null hypothesis that the measure of association is 0.

Exact two-sided *p* values (possibly up to Monte Carlo accuracy) for testing the null hypothesis that the measure of association is 0.

Representing Data in Crosstabular Form

All of the measures of association considered in this book are defined from data that can be represented in the form of the *r c* contingency table, as shown in Table 13.1.

This table is formed from N observations cross-classified into row categories (r) and column categories (c), with of the observations falling into row category i and column category j. Such a table is appropriate for categorical data. For example, the row classification might consist of three discrete age categories (young, middle-aged, and elderly), and the column classification might consist of three discrete annual income categories ($25,000–50,000, $50,000–75000, and $75,000–100,000). These are examples of ordered categories. Alternatively, one or both of the discrete categories might be nominal. For example, the row classification might consist of three cities (Boston, New York, and Philadelphia). In this chapter, you will define various measures of association based on crosstabulations such as the one shown in Table 13.1.

Measures of association are also defined on data sets generated from continuous bivariate distributions. Although such data sets are not naturally represented as crosstabulations, it is neverthe

Point Estimates

Maximum-likelihood theory is used to estimate each meas

tion. The exact two-sided *p* value is obtained by Equation 9.4, with substituted for . Thus,

Equation 13.2

An equivalent definition of the two-sided *p* value is

Equation 13.3

This definition expresses the exact two-sided *p* value as a sum of two exact one-sided *p* values, one in the left tail and the other in the right tail of the exact distribution of . Exact permutational distributions are not usually symmetric, so the areas in the two tails may not be equal. This is an important distinction between exact and asymptotic *p* values. In the latter case, the exact two-sided *p* value is always double the exact one-sided *p* value by the symmetry of the asymptotic normal distribution of .

Monte Carlo P Values

Monte Carlo *p* values are very close approximations to corresponding exact *p* values but have the advantage that they are much easier to compute. These *p* values are computed by the methods described in Chapter 9 in "Monte Carlo Two-Sided P Values" on p. 143. For nominal data, only two-sided *p* values are defined. The Monte Carlo estimate of the exact two-sided *p* value is obtained by Equation 9.6, with an associated confidence interval given by Equation 9.8. In this computation, the critical region is defined by

Equation 13.4

For measures of association based on ordinal data and for measures of agreement, two-sided *p* values are defined. For two-sided *p* values,

Equation 13.5

Asymptotic P Values

For measures of association based on nominal data, only two-sided *p* values are defined.

For measures of association on ordinal data and for measures of agreement, the asymptotic standard error of the maximum-likelihood estimate under the null hypothesis (ASE0) is obtained. Then asymptotic one- and two-sided *p* values are obtained by using the fact that the ratio $M(x)/ASEO$ converges to a standard normal distribution.

14 Measures of Association for Ordinal Data

Exact Tests provides the following measures of association between pairs of ordinal variables: Pearson's product-moment correlation coefficient, Spearman's rank-order correlation coefficient, Kendall's tau coefficient, Somers' *d* coefficient, and the gamma coefficient. All of these measures of association range between –1 and +1, with 0 signifying no association, –1 signifying perfect negative association, and +1 signifying perfect positive association. One other measure of association mentioned in this chapter is Kendall's *W*, also known as Kendall's coefficient of concordance. This test is discussed in detail in Chapter 7.

Available Measures

Table 14.1 shows the available measures of association, the procedure from which each can be obtained, and a bibliographical reference for each test.

Table 14.1 Available tests

Measure of Association	**Procedure**	**Reference**
Pearson's product-moment correlation	Crosstabs	Siegel and Castellan (1988)
Spearman's rank-order correlation	Crosstabs	Siegel and Castellan (1988)
Kendall's *W*	Nonparametric Tests: Tests for Several Related Samples	Conover (1975)
Kendall's tau-*b*, Kendall's tau-*c*, and Somers' *d*	Crosstabs	Siegel and Castellan (1988)
Gamma coefficient	Crosstabs	Siegel and Castellan (1988)

Pearson's Product-Moment Correlation Coefficient

Let *A* and *B* be a pair of correlated random variables. Suppose you observe *N* pairs of observations and crosstabulate them into the contingency table displayed as Table 13.1, in which the 's are the distinct values assumed by *A* and the 's are the distinct values assumed by *B*. When the data follow a bivariate normal distribution, the appropriate measure of association is the correlation coefficient, , between *A* and *B*. This parameter is estimated by Pearson's product-moment correlation coefficient, shown in Equation 14.1. In this equation, represents the marginal row total and represents the marginal column total.

Equation 14.1

where

Equation 14.2

The formulas for the asymptotic standard errors are fairly complicated. These formulas are discussed in the algorithms manual available on the Manuals CD and also available by selecting

Figure 14.2 Pearson's product-moment correlation coefficient for subset of social status striving data

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Exact Significance
Interval by Interval	Pearson's R	.739	.054	2.452	.058[1]	.037
N of Valid Cases		7				

1. Based on normal approximation

The correlation coefficient has a point estimate of $R = 0.739$. The exact two-sided p value is 0.037 and indicates that the correlation coefficient is significantly different from 0. The corresponding asymptotic two-sided p value is 0.058 and fails to demonstrate statistical significance at the 5% level for this small data set.

It should be noted that the computational limits for exact inference are reached rather quickly for Pearson's product-moment correlation coefficient with continuous data. By the time $N = 10$, the Monte Carlo option should be used rather than the exact option. Consider, for example, the complete authoritarianism data set of 12 observations (Siegel and Castellan, 1988) shown in Figure 14.3.

Figure 14.3 Complete social status striving data

For this data set, the exact two-sided p value, shown in Figure 14.5, is 0.001, approximately half the asymptotic two-sided p value of 0.003. However, it may be time-consuming to perform the exact calculation. In contrast, the Monte Carlo p value based on 10,000 samples from the data set produces a significance estimate of 0.002, practically the same as the exact p value. The 99% confidence interval for the exact p

value is (0.001, 0.003). The Monte Carlo output is shown in Figure 14.4, and the

for . Once these transformations are made, all of the remaining calculations for the point estimate (R), the standard error (ASE1), the confidence interval, the asymptotic p value, and the exact p value are identical to corresponding ones for Pearson's product-moment correlation coefficient.

Consider, for example, the data displayed in Figure 13.1. Figure 14.6 displays these

defines the number of pairs of observations that are concordant relative to the observations in cell (i, j), and the formula

$$D_{ij} = \sum_{h<i\ k>j} x_{hk} + \sum_{h>i\ k<j} x_{hk} \qquad \text{Equation 14.7}$$

defines the number of pairs of observations that are discordant relative to the observations in cell (i, j). Thus, the total number of concordant pairs in the entire data set is

$$P = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} C_{ij} \qquad \text{Equation 14.8}$$

and the total number of discordant pairs in the entire data set is

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} D_{ij} \qquad \text{Equation 14.9}$$

Kendall's tau and Somers' d and their various variants are functions of $P - Q$. Thus, although their respective point estimates and standard errors differ, they all produce the same p values. Next, these measures of association will be defined and their use illustrated through a numerical example.

Kendall's Tau-b and Kendall's Tau-c

Kendall's tau coefficient has three variants, , $_b$, and $_c$. You first specify estimators and associated asymptotic standard errors for these three variants. For a discussion of the criteria for selecting one variant over another, see Gibbons (1993). The $_b$ and $_c$ variants were developed to correct for ties and for categorical data.

Kendall's $_b$ coefficient is estimated by

$$T_b = \frac{P - Q}{\sqrt{D_r D_c}} \qquad \text{Equation 14.10}$$

where

$$D_r = N^2 - \sum_{i=1}^{r} m_i \qquad \text{Equation 14.11}$$

and

Equation 14.12

Kendall's coefficient is estimated by

Equation 14.13

where .

Somers' d

Somers' *d* coefficient is a useful measure of association between two asymmetrically related ordinal variables, where one of the two variables is regarded as independent and the other as dependent. See Siegel and Castellan (1988) for a discussion of this

Example: Smoking Habit Data

Observe that all variants of Kendall's tau and Somers' *d* are functions of . They differ only in how they are standardized. Thus, although their point estimates and asymptotic standard errors vary, the exact and asymptotic *p* values for testing the null hypothesis that there is no association are invariant across all these measures. Cnl1s

Although all of these coefficients have different point estimates, their sampling distributions are equivalent, thus leading to a common p value. The exact two-sided p value for testing the null hypothesis that there is no association is 0.0226, and the corresponding asymptotic two-sided p value is 0.0177.

As the number of observations grows, it becomes increasingly difficult to compute exact p

Figure 14.13 shows the Monte Carlo results for the full data set. The Monte Carlo sample size was 10,000.

Figure 14.13 Monte Carlo results for Kendall's tau and Somers' d for full smoking data

Directional Measures

.338	.046	7.339	.000	.000[3]	.000	.000
.282	.038	7.339	.000	.000[3]	.000	.000
.420	.05362.46 46sg[(.)6(re)3s13(6(refBT6.-8S31 9631 0 0 6.9631 264.78 457.8 Tm0 g[(7					

It is clear that a strong correlation exists between the duration and status of the smoking habit. The exact two-sided *p* value for testing the null hypothesis that there is no correlation is at most 0.0003 with 95% confidence.

Gamma Coefficient

The gamma coefficient is yet another measure of association between two ordinal variables. It was first discussed extensively by Goodman and Kruskal (1963). It is an alternative to Kendall's tau and Somers' *d* for ordered categorical variables. Like these measures, it is defined in terms of the difference between concordant and discordant pairs, and so does not require the variables to take on actual numerical values. Using the notation developed in the previous section, the gamma coefficient is estimated by

$$G = \frac{P - Q}{P + Q} \qquad \text{Equation 14.17}$$

If the data contain no ties, this definition of gamma will yield the same exact and asymptotic *p* values as Kendall's tau and Somers' *d*. In general, however, inference based on gamma can differ from inference based on the latter two coefficients. You can now analyze the small data set of cessation and smoking habit displayed in Figure 14.10. Figure 14.14 displays point and interval estimates of gamma along with exact and asymptotic *p* values for testing the null hypothesis that there is no association.

Figure 14.14 Gamma coefficient for subset of smoking data

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Exact Significance
Ordinal by Ordinal	Gamma	.345	.140	2.372	.018	.024
N of Valid Cases		96				

The gamma coefficient is estimated as 0.345. The exact two-sided *p* value for testing the null hypothesis that there is no association is 0.024.

As the number of observations grows, it becomes increasingly difficult to compute exact *p* values, and the Monte Carlo option is a better choice. Figure 14.15 shows the Monte Carlo results for the full cessation and smoking habit data set shown in Figure 14.12. The Monte Carlo sample size was 10,000.

Figure 14.15 Monte Carlo results for gamma coefficient for full smoking data

Symmetric Measures

		Value	Asymp. Std. Error[1]	Approx. T[2]	Approx. Sig.	Monte Carlo Sig.		
						Sig.	99% Confidence Interval	
							Lower Bound	Upper Bound
Ordinal by Ordinal	Gamma	.483	.064	7.339	.000	.000[3]	.000	.000
N of Valid Cases		240						

1. Not assuming the null hypothesis.
2. Using the asymptotic standard error assuming the null hypothesis.
3. Based on 10000 sampled tables with starting seed 2000000.

It is clear that a strong correlation exists between the duration and status of the smoking habit. The exact two-sided p value for testing the null hypothesis that there is no correlation is at most 0.0005 with 99% confidence.

of these measures have an identical two-sided *p* value for testing the null hypothesis that there is no association, which is the same as the Pearson chi-square *p* value and which is based on the distribution of . Exact Tests reports both the asymptotic and exact *p* values.

The formulas for computing the three contingency coefficients are given below. The formula for each measure involves taking the square root of a function of . The positive root is always selected. For a more detailed discussion of these measures of association, see Liebetrau (1983).

The phi contingency coefficient is given by the formula

Equation 15.1

The minimum value assumed by is 0, signifying no association. However, its upper bound is not fixed but depends on the dimensions of the contingency table. Therefore, it is not a very suitable measure for arbitrary tables. For the special case of the table, Gibbons (1985) shows that is identical to the absolute value of Kendall's coefficient and is evaluated by the formula

Equation 15.2

Notice from Equation 15.2 that, for the contingency table, could be either positive or negative, which implies a positive or negative association in the table.

The Pearson contingency coefficient is given by the formula

Equation 15.3

This contingency coefficient assumes a minimum value of 0, signifying no association. It is bounded from above by 1, signifying perfect association. However, the maximum value attainable by *CC* is , where . Thus, the range of this contingency coefficient still depends on the dimensions of the table. Cramér's *V* coefficient ranges between 0 and 1, with 0 signifying no association and 1 signifying perfect association. It is given by

Equation 15.4

Exact Tests reports the point estimate of the contingency coefficient. The formulas for these asymptotic standard errors are fairly complicated. These formulas are described in the algorithms manual available on the Manuals CD and also available by selecting Algorithms

These measures may be used to analyze an unordered contingency table given in Siegel and Castellan (1988). The data consist of a crosstabulation of three possible responses (*completed*, *declined*, *no response*) to a questionnaire concerning the financial accounting standards used by six different organizations responsible for maintaining such standards. These organizations are identified only by their initials (*AAA, AICPA, FAF, FASB, FEI,* and *NAA*). The crosstabulated data are shown in Figure 15.1.

Figure 15.1 Crosstabulation of response to survey and finance organization

Count

8	8	3	11	17	2
2	5	1	2		

First, these data are analyzed using only the first three columns of Figure 15.1. For this subset of the data, Figure 15.2 shows the results for the contingency coefficients. The exact two-sided *p* value for testing the null hypothesis that there is no association is also reported. Its value is 0.090, slightly lower than the asymptotic *p* value of 0.092.

The next analysis uses the full data set, which consists of all six columns of Figure 15.1. This data set is too large to compute the exact *p* value. However, a 99% confidence interval on the exact *p* value based on 10,000 Monte Carlo samples is easily obtained. The results are shown in Figure 15.3.

Figure 15.3 Monte Carlo results for phi and Cramér's V

Symmetric Measures

		Value	Approx. Sig.	Monte Carlo Significance		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Nominal by Nominal	Phi	.723	.000	.0000[1]	.0000	.0005
	Cramer's V	.511	.000	.0000[1]	.0000	.0005
N of Valid Cases		144				

1. Based on 10000 and seed 2000000 ...

The p value for testing the null hypothesis that there is no association is at most 0.0005 with 99% confidence, which implies that the row and column classifications are not independent.

Proportional Reduction in Prediction Error

In regression problems involving continuous data, the coefficient of determination (or R^2 statistic) is often used to measure the proportion of the total variation attributable to the explanatory variable. It would be useful to provide an analog of this index for nominal categorical data. Two measures of association are available for this purpose. One is Goodman and Kruskal's tau, and the other is the uncertainty coefficient. Both measure the proportion of variation in the row variable that can be attributed to the column variable.

Goodman and Kruskal's Tau

Goodman and Kruskal's tau coefficient for measuring the proportion of the variation in the row variable attributable to the column variable is estimated by

$$\hat{}_{R|C}(x) = \frac{\Sigma_{j=1}^{c} n \Sigma_j^{-1} \Sigma_{i=1}^{r} x_{ij}^2 - N^{-1} \Sigma_{i=1}^{r} m_i}{N - N^{-1} \Sigma_{i=1}^{r} m_i^2} \qquad \text{Equation 15.5}$$

This coefficient ranges between 0 and 1, with 0 implying no reduction in row variance when the column category is known, and 1 imp

Count

		Preferred Cold War	
		U.S.	U.S.S.R.
Party Preference	Right	225	3
	Center	53	1
	Left	206	12

First, Goodman and Kruskal's tau is estimated, a confidence interval is obtained for it, and the null hypothesis that there is no association in the population is tested. The results are shown in Figure 15.5.

The observed value of Goodman and Kruskal's tau with ally, 0.013, is rather small and leads to the conclusion that 1.3% of the variation in choice of preferred ally is explained by knowing a person's party preference. The exact *p* value, 0.045, implies that the null hypothesis that there is no association can be rejected at the 5% level. In other words, the small amount of explained variation is real, not due to sampling error.

Next, the uncertainty coefficient is estimated, a confidence interval is obtained for it, and the null hypothesis that there is no association in the population is tested. The results are shown in Figure 15.6.

Figure 15.6 Uncertainty coefficient for party preference and preferred cold war ally data

			Value	Asymp. Std. Error[1]	Approx. T[2]	Approx. Sig.	Exact Significance
Nominal by Nominal	Uncertainty Coefficient	Symmetric	.012	.009	1.346	.033[3]	.034
		Party Preference Dependent	.007	.005	1.346	.033[3]	.034
		Preferred Cold War Ally Dependent	.048	.034	1.346	.033[3]	.034

Once again, the observed value of the uncertainty coefficient with ally, 0.007, is extremely small. However, the exact two-sided *p* value, 0.034, is statistically significant and indicates that the measure is indeed greater than 0.

Measures of Agreement

This chapter discusses kappa, a measure used to assess the level of agreement between two observers classifying a sample of objects on the same categorical scale. The joint ratings of the observers are displayed on a square contingency table such as Table 13.1. Kappa (see Agresti, 1990) can be obtained using the Crosstabs procedure.

Kappa

The kappa coefficient is defined on a square contingency table. It is estimated by

Equation 16.1

Notice that the kappa statistic does not depend on the off-diagonal elements of the

Figure 16.1 Crosstabulation of student teachers rated by supervisors (partial data)

The results for the kappa statistic are shown in Figure 16.2.

The value of kappa is estimated at . The positive sign on the kappa statistic implies that the agreement is positive. The exact two-sided *p* value of 0.048 is significant; thus, you can reject the null hypothesis that there is no agreement. Notice, however, that the asymptotic two-sided *p* value is not very accurate for this small data set. It is less than one half of the exact *p* value.

The same analysis conducted with the full data set of 72 observations is tabulated in Figure 16.3.

For this larger data set, it is more efficient to perform the Monte Carlo inference rather than the exact inference. Figure 16.4 shows the results based on 10,000 Monte Carlo samples.

Figure 16.4 Monte Carlo results for student teacher ratings data

.362	.091	4.329	.000	.0000[1]	.0000	.0005
72						

In the full data set, the kappa statistic has a smaller value, 0.362. However, due to the larger sample size this observed statistic is highly significant, with a two-sided *p* value guaranteed to be less than 0.0005 with 99% confidence.

Syntax Reference

Exact Tests Syntax

The METHOD subcommand allows you to specify the method used to calculate significance levels. The MH subcommand performs the marginal homogeneity test. The J-T subcommand performs the Jonckheere-Terpstra test. See the *Syntax Reference Guide* for a complete description of the full

5 minutes. If a test exceeds a time limit of 30 minutes, it is recommended that you use the Monte Carlo, rather than the exact, method.

MH Subcommand

```
NPAR TESTS /MH=varlist [WITH varlist [(PAIRED)]]
```

MH performs the marginal homogeneity test, which tests whether combinations of values between two paired ordinal variables are equally likely. The marginal homogeneity test is typically used in repeated measures situations. This test is an extension of the McNemar test from binary response to multinomial response. The output shows the number of distinct values for all test variables, the number of valid off-diagonal cell counts, mean, standard deviation, observed and standardized values of the test statistics, the asymptotic two-tailed probability for each pair of variables, and, if a /METHOD

This example performs the marginal homogeneity test on variable pairs *V1* and *V2*, *V1* and *V3*, and *V2* and *V3*. The exact *p* values are estimated using the Monte Carlo sampling method.

J-T Subcommand

```
NPAR TESTS /J-T=varlist BY variable(value1,value2)
```

J-T (alias JONCKHEERE-TERPSTRA) performs the Jonckheere-Terpstra test, which tests whether *k* independent samples defined by a grouping variable are from the same population. This test is particularly powerful when the *k* populations have a natural ordering. The output shows the number of levels in the grouping variable, the total number of cases, observed, standardized, mean and standard deviation of the test statistic, the two-tailed asymptotic significance, and, if a `/METHOD` subcommand is specified, one-tailed and two-tailed exact or Monte Carlo probabilities.

Syntax

The minimum specification is a test variable, the keyword BY, a grouping variable, and a pair of values in parentheses.

Every value in the range defined by the pair of values for the grouping variable forms a group.

If the `/METHOD` subcommand is specified, and the number of populations, *k*, is greater than 5, the *p* value is estimated using the Monte Carlo sampling method. The exact *p* value is not available when *k* exceeds 5.

Operations

Cases from the *k* groups are ranked in a single series, and the rank sum for each group is computed. A test statistic is calculated for each variable specified before BY.

The Jonckheere-Terpstra statistic has approximately a normal distribution.

Cases with values other than those in the range specified for the grouping variable are excluded.

The direction of a one-tailed inference is indicated by the sign of the standardized test statistic.

Example

```
NPAR TESTS /J-T=V1 BY V2(0,4)
  /METHOD=EXACT.
```

This example performs the Jonckheere-Terpstra test for groups defined by values 0 through 4 of *V2*. The exact *p* values are calculated.

Appendix B
Algorithms in Exact Tests

Exact Algorithms

An exact *p* value is computed by enumerating every single outcome in some suitably defined reference set, identifying all outcomes that are more extreme than the observed one, and summing their probabilities under the null hypothesis. Although this might appear to be a formidable computing problem by the time the size of the reference set exceeds, say, a few million, it is still feasible. Many researchers have worked on this problem and have developed fast numerical algorithms that enumerate all of the possible outcomes *implicitly* rather than *explicitly.* That is, these algorithms don't examine each individual outcome separately. There are ways to identify large numbers of outcomes at one time and classify them as either more or less extreme than the observed outcome. A complete collection of reference files for all of these algorithms is available in the Exact-Stats Mailbase on the Internet. These references can be accessed through FTP, Gopher, or World Wide Web at the following addresses:

```
ftp://mailbase.ac.uk/pub/lists/exact-stats/files
```

```
gopher://mailbase.ac.uk/Mailbase Lists - A-E/exact-stats/Other Files
```

```
http://www.mailbase.ac.uk/Mailbase Lists - A-E/exact-stats/Other Files
```

One class of algorithms, called network algorithms, was developed by Mehta, Patel, and their colleagues at the Harvard School of Public Health. These algorithms are referenced below in chronological order. Many of them have already been incorporated into Exact Tests, and others will be incorporated into future releases of the software.

Mehta, C. R., and N. R. Patel. 1980. A netw

Mehta, C. R., N. R. Patel, and R. Gray. 1985. On computing an exact confidence interval for the common odds ratio in several 2 2 contingency tables. *Journal of the American Statistical Association*, 80:392, 969–973.

Mehta, C. R., and N. R. Patel. 1986. A hybrid algorithm for Fisher's exact test in unordered *r c* contingency tables. *Communications in Statistics*, 15:2, 387–403.

Mehta, C. R., and N. R. Patel. 1986. FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered *r c* contingency tables. *ACM Transactions on Mathematical Software*, 12:2, 154–161.

Hirji, K., C. R. Mehta, and N. R. Patel. 1987. Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, 82:400, 1110–1117.

Mehta, C. R., N. R. Patel, and L. J. Wei. 1988. Constructing exact significance tests with restricted randomization rules. *Biometrika*, 75:2, 295–302.

Hirji, K., C. R. Mehta, and N. R. Patel. 1988. Exact inference for matched case control studies. *Biometrics*, 44:3, 803–814.

Agresti, A., C. R. Mehta, and N. R. Patel. 1990. Exact inference for contingency tables with ordered categories. *Journal of the American Statistical Association*, 85:410, 453–458.

Mehta, C. R., N. R. Patel, and P. Senchaudhuri. 1992. Exact stratified linear rank tests for ordered categorical and binary data. *Journal of Computational and Graphical Statistics*, 1: 21–40.

Mehta, C. R. 1992. An interdisciplinary approach to exact inference for contingency tables. *Statistical Science*, 7: 167–170.

Hilton, J., and C. R. Mehta. 1993. Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics*, 49: 609–616.

Hilton, J., C. R. Mehta, and N. R. Patel. 1994. Exact Smirnov *p* values using a network algorithm. *Computational Statistics and Data Analysis*, 17:4, 351–361.

Mehta, C. R., N. R. Patel, P. Senchaudhuri, and A. A. Tsiatis. 1994. Exact permutational tests for group sequential clinical trials. *Biometrics*, 50:4, 1042–1053.

Monte Carlo Algorithms

Monte Carlo algorithms solve a slightly easier computational problem. They do not attempt to enumerate all of the members of the reference set. Instead, they estimate the *p* value by taking a random sample from the reference set. The Monte Carlo algorithms in Exact Tests make use of ideas in the following papers (in chronological order):

Agresti, A., D. Wackerly, and J. M. Boyett. 1979. Exact conditional tests for cross-classifications: Approximations of attained significance levels. *Psychometrika*, 44: 75–83.

Patefield, W. M. 1981. An efficient method of generating *r c* tables with given row and column totals. (Algorithm AS 159.) *Applied Statistics*, 30: 91–97.

Mehta, C. R., N. R. Patel, and P. Senchaudhuri. 1988. Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association*, 83:404, 999–1005.

Senchaudhuri, P., C. R. Mehta, and N. R. Patel. 1995. Estimating exact *p* values by the method of control variates, or Monte Carlo rescue. *Journal of American Statistical Association.*

Appendix C
Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

10504-1785

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and canno

Bibliography

Agresti, A. 1990. *Categorical data analysis*. New York: John Wiley and Sons.

_____. 1992. A survey of exact inference for contingency tables. *Statistical Science*, 7:1, 131–177.

Agresti, A., and M. C. Yang. 1987. An empirical investigation of some effects of sparseness in contingency tables.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 5*, 50: 157–175.

Pitman, E. J. G. 1948. Notes on non-parametric statistical inference. Columbia University (duplicated).

Pratt, J. W., and J. D. Gibbons. 1981. *Concepts of nonparametric theory*. New York: Springer-Verlag.

Radlow, R., and E. F. Alf. 1975. An alternate multinomial assessment of the accuracy of the chi-square test of goodness of fit. *Journal of the American Statistical Association*, 70: 811–813.

Read, T. R., and N. A. Cressie. 1988. *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer-Verlag.

Roscoe, J. T., and J. A. Byars. 1971. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association*, 66:336, 755–759.

Senchaudhuri, P., C. R. Mehta, and N. R. Patel. 1995. Estimating exact *p* values by the method of control variates, or Monte Carlo rescue. *Journal of the American Statistical Association* (forthcoming).

Siegel, S. 1956. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Siegel, S., and N. J. Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. 2nd ed. New York: McGraw-Hill.

Smirnov, N. V. 1939. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, 2:2, 3–16.

Snapinn, S. M., and R. D. Small. 1986. Tests of significance using regression models for ordered categorical data. *Biometrics*, 42: 583–592.

Sprent, P. 1993. *Applied nonparametric statistical methods*. 2nd ed. London: Chapman and Hall.

Wald, A., and J. Wolfowitz. 1940. On a test whether two samples are from the same population. *Annals n 1 2 Tw[y(6ls4:))land 2FYg4 6TJTT4 1 Tf4.3418 TD-0 0 T11: 14(,)w(, 42:(fo3 0TJ-21.31*

Index