Exact Tests™

Cyrus R. Mehta and Nitin R. Patel

For more information about SPSS® software products, please visit our WWW site at *http://www.spss.com* or contact

Marketing Department
SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6307
Tel: (312) 651-3000
Fax: (312) 651-3668

SPSS is a registered trademark.
PASW is a registered trademark of SPSS Inc.

The SOFTWARE and documentation are provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subdivision (c)(1)(ii) of The Rights in Technical Data and Computer Software clause at 52.227-7013. Contractor/manufacturer is SPSS Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL, 60606-6307.

General notice: Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

TableLook is a trademark of SPSS Inc.
Windows is a registered trademark of Microsoft Corporation.
Printed in the United States of America.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

Preface

Exact Tests is a statistical package for analyzing continuous or categorical data by exact methods. The goal in Exact Tests is to enable you to make reliable inferences when your data are small, sparse, heavily tied, or unbalanced and the validity of the corresponding large sample theory is in doubt. This is achieved by computing exact *p* values for a very wide class of hypothesis tests, including one-, two-, and K- sample tests, tests for unordered and ordered categorical data, and tests for measures of association. The statistical methodology underlying these exact tests is well established in the statistical literature and may be regarded as a natural generalization of Fisher's exact test for the single 2×2 contingency table. It is fully explained in this user manual. The real challenge has been to make this methodology operational through software development. Historically, this has been a difficult task because the computational demands imposed by the exact methods are rather severe. We and our colleagues at the Harvard School of Public Health have worked on these computational problems for over a decade and have developed exact and Monte Carlo algorithms to solve them. These algorithms have now been implemented in Exact Tests. For small data sets, the algorithms ensure quick computation of exact *p* values. If a data set is too large for the exact algorithms, Monte Carlo algorithms are substituted in their place in order to estimate the exact *p* values to any desired level of accuracy.

These numerical algorithms are fully integrated into the PASW Statistics system. Simple selections in the Nonparametric Tests and Crosstabs dialog boxes allow you to obtain exact and Monte Carlo results quickly and easily.

Acknowledgments

Exact Tests is the result of a collaboration between Cytel Software Corporation and SPSS Inc. The exact algorithms were developed by Cytel. Integrating the exact engines into the user interface and documenting the statistical methods in a comprehensive user manual were tasks shared by both organizations. We would like to thank our fellow developers, Yogesh Gajjar, Hemant Govil, Pralay Senchaudhuri, and Shailesh Vasundhara of Cytel.

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 1

The Exact Method 1
The Monte Carlo Method 3
When to Use Exact Tests 4
How to Obtain Exact Statistics 6
Additional Features Available with Command Syntax 8
Nonparametric Tests 8
How to Set the Random Number Seed 8
Pivot Table Output 9

2 Exact Tests 11

Pearson Chi-Square Test for a 3 x 4 Table 14
Fisher's Exact Test for a 2 x 2 Table 18
Choosing between Exact, Monte Carlo, and Asymptotic P Values 22
- When to Use Exact P Values 24
- When to Use Monte Carlo P Values 24
- When to Use Asymptotic P Values 29

3 One-Sample Goodness-of-Fit Inference 39

Available Tests 39
Chi-Square Goodness-of-Fit Test 39
- Example: A Small Data Set 42
- Example: A Medium-Sized Data Set 44

One-Sample Kolmogorov Goodness-of-Fit Test 45
- Example: Testing for a Uniform Distribution 47

4 One-Sample Inference for Binary Data 49

Available Tests 49

Binomial Test and Confidence Interval 49

Example: Pilot Study for a New Drug 50

Runs Test 51

Example: Children's Aggression Scores 53

Example: Small Data Set 54

5 Two-Sample Inference: Paired Samples 57

Available Tests 57

When to Use Each Test 58

Statistical Methods 59

Sign Test and Wilcoxon Signed-Ranks Test 59

Example: AZT for AIDS 64

McNemar Test 68

Example: Voters' Preference 70

Marginal Homogeneity Test 71

Example: Matched Case-Control Study of Endometrial Cancer 71

Example: Pap-Smear Classification by Two Pathologists 72

6 Two-Sample Inference: Independent Samples 75

Available Tests 75

When to Use Each Test 76

Statistical Methods 76

The Null Distribution of T 79

P Value Calculations 80

Mann-Whitney Test 80

Exact P Values 82

Monte Carlo P Values 83

Asymptotic P Values 84

Example: Blood Pressure Data 84

Kolmogorov-Smirnov Test 87

Example: Effectiveness of Vitamin C 90

Wald-Wolfowitz Runs Test 91
Example: Discrimination against Female Clerical Workers 92
Median Test 94

7 K-Sample Inference: Related Samples 95

Available Tests 95
When to Use Each Test 96
Statistical Methods 96
Friedman's Test 101
Example: Effect of Hypnosis on Skin Potential 102
Kendall's W 104
Example: Attendance at an Annual Meeting 105
Example: Relationship of Kendall's W to Spearman's R 107
Cochran's Q Test 108
Example: Crossover Clinical Trial of Analgesic Efficacy 109

8 K-Sample Inference: Independent Samples 113

Available Tests 113
When to Use Each Test 114
Tests Against Unordered Alternatives 114
Tests Against Ordered Alternatives 115
Statistical Methods 116
Distribution of T 119
P Value Calculations 119
Median Test 122
Example: Hematologic Toxicity Data 125
Kruskal-Wallis Test 127
Example: Hematologic Toxicity Data, Revisited 129
Jonckheere-Terpstra Test 131
Example: Space-Shuttle O-Ring Incidents Data 132

9 Introduction to Tests on R x C Contingency Tables 135

Defining the Reference Set 137

Defining the Test Statistic 138

Exact Two-Sided P Values 138

Monte Carlo Two-Sided P Values 139

Asymptotic Two-Sided P Values 140

10 Unordered R x C Contingency Tables 141

Available Tests 141

When to Use Each Test 141

Statistical Methods 142

Oral Lesions Data 143

Pearson Chi-Square Test 144

Likelihood-Ratio Test 145

Fisher's Exact Test 147

11 Singly Ordered R x C Contingency Tables 149

Available Test 149

When to Use the Kruskal-Wallis Test 149

Statistical Methods 149

Tumor Regression Rates Data 150

12 Doubly Ordered R x C Contingency Tables 155

Available Tests 155

When to Use Each Test 156

Statistical Methods 156

Dose-Response Data 157

Jonckheere-Terpstra Test 158

Linear-by-Linear Association Test 161

13 Measures of Association 165

Representing Data in Crosstabular Form 165

Point Estimates 168

Exact P Values 168

- Nominal Data 168
- Ordinal and Agreement Data 168

Monte Carlo P Values 169

Asymptotic P Values 169

14 Measures of Association for Ordinal Data 171

Available Measures 171

Pearson's Product-Moment Correlation Coefficient 172

- Spearman's Rank-Order Correlation Coefficient 174

Kendall's W 177

Kendall's Tau and Somers' d Coefficients 177

- Kendall's Tau-b and Kendall's Tau-c 178
- Somers' d 179
- Example: Smoking Habit Data 180

Gamma Coefficient 183

15 Measures of Association for Nominal Data 185

Available Measures 185

Contingency Coefficients 185

Proportional Reduction in Prediction Error 188

- Goodman and Kruskal's Tau 188
- Uncertainty Coefficient 189
- Example: Party Preference Data 189

16 Measures of Agreement 193

Kappa 193

- Example: Student Teacher Ratings 193

CROSSTABS 199
Exact Tests Syntax 199
METHOD Subcommand 199
NPAR TESTS 200
Exact Tests Syntax 200
METHOD Subcommand 200
MH Subcommand 201
J-T Subcommand 202

Appendix A
Conditions for Exact Tests 203

Appendix B
Algorithms in Exact Tests 205

Exact Algorithms 205

Monte Carlo Algorithms 206

Bibliography 207

Index 211

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 fail to meet any of the underlying assumptions necessary for reliable results using the standard asymptotic method.

The Exact Method

By default, PASW Statistics calculates significance levels for the statistics in the Crosstabs and Nonparametric Tests procedures using the **asymptotic method**. This means that *p* values are estimated based on the assumption that the data, given a sufficiently large sample size, conform to a particular distribution. However, when the data set is small, sparse, contains many ties, is unbalanced, or is poorly distributed, the asymptotic method may fail to produce reliable results. In these situations, it is preferable to calculate a significance level based on the exact distribution of the test statistic. This enables you to obtain an accurate *p* value without relying on assumptions that may not be met by your data.

The following example demonstrates the necessity of calculating the *p* value for small data sets. This example is discussed in detail in Chapter 2.

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.

Figure 1.3 Exact results of Pearson chi-square test for fire fighter data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)
Pearson Chi-Square	11.556[1]	6	.073	.040

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

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 Table 1.1 and Table 1.2 for details.) The Monte Carlo method is a repeated sampling method. For any observed table, there are many tables, each with the same dimensions and column and row margins as the observed table. The Monte Carlo method repeatedly samples a spec-

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

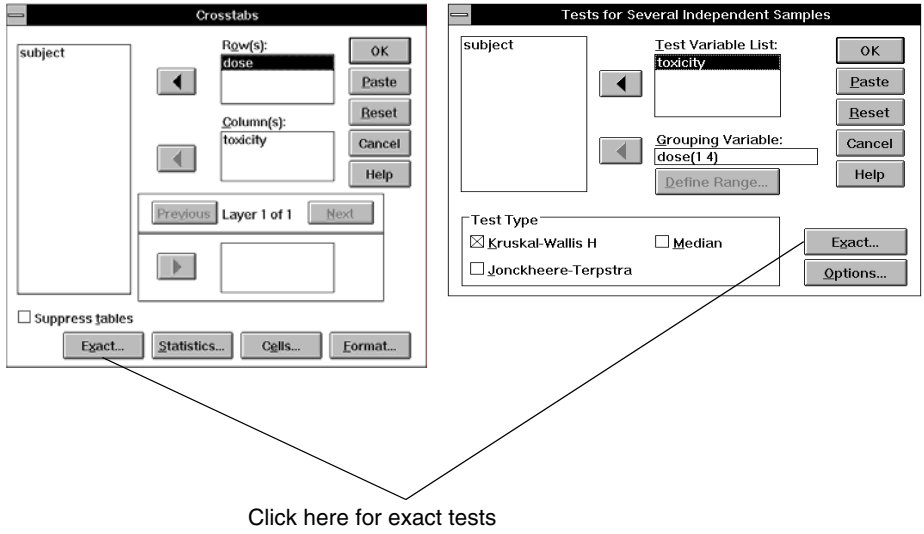
2 x 2 contingency tables (obtained by selecting chi-square)	
Pearson chi-square test	$N \leq 100,000$
Fisher's exact test	$N \leq 100,000$
Likelihood-ratio test	$N \leq 100,000$
r x c contingency tables (obtained by selecting chi-square)	
Pearson chi-square test	$N \leq 30$ and $\min\{r, c\} \leq 3$
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.

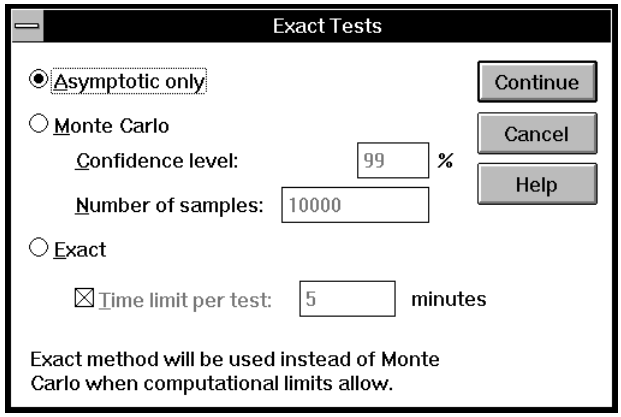
To obtain exact statistics, open the Crosstabs dialog box or any of the Nonparametric Tests dialog boxes. The Crosstabs and Tests for Several Independent Samples dialog boxes are shown in Figure 1.5.

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.

Figure 1.6 Exact Tests dialog box



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 calculated quickly for your data, they will be provided. See Appendix A for details on the circumstances under which exact, rather than Monte Carlo, results are provided. Note that, within a session, the Monte Carlo method relies on a random number seed that changes each time you run the procedure. If you want to duplicate your results, you should set the random number seed every time you use the Monte Carlo method. See "How to Set the Random Number Seed" on p. 9 for more information.

Confidence level. Specify a confidence level between 0.01 and 99.9. The default value is 99.

Number of samples. Specify a number between 1 and 1,000,000,000 for the number of samples used in calculating the Monte Carlo approximation. The default is 10,000. Larger numbers of samples produce more reliable estimates of the exact *p* value but also take longer to calculate.

Exact. Calculates the exact *p* value. Asymptotic results are also displayed. Because computing exact statistics can be time-consuming, you can set a limit on the amount of time allowed for each test.

Time limit per test. Enter the maximum time allowed for calculating each test. The time limit can be between 1 and 9,999,999 minutes. The default is five minutes. If the time limit is reached, the test is terminated, no exact results are provided, and the application proceeds to the next test in the analysis. If a test exceeds a set time limit of 30 minutes, it is recommended that you use the Monte Carlo, rather than the exact, method.

Calculating the exact *p* value can be memory-intensive. If you have selected the exact method and find that you have insufficient memory to calculate results, you should first close any other applications that are currently running in order to make more memory available. If you still cannot obtain exact results, use the Monte Carlo method.

Additional Features Available with Command Syntax

Command syntax allows you to:

- Exceed the upper time limit available through the dialog box.
- Exceed the maximum number of samples available through the dialog box.
- Specify values for the confidence interval with greater precision.

Nonparametric Tests

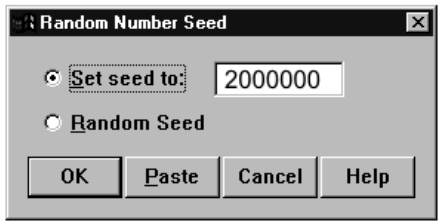
As of release 6.1, two new nonparametric tests became available, the Jonckheere-Terpstra test and the marginal homogeneity test. The Jonckheere-Terpstra test can be obtained from the Tests for Several Independent Samples dialog box, and the marginal homogeneity test can be obtained from the Two-Related-Samples Tests dialog box.

How to Set the Random Number Seed

Monte Carlo computations use the pseudo-random number generator, which begins with a **seed**, a very large integer value. Within a session, the application uses a different seed each time you generate a set of random numbers, producing different results. If you want to duplicate your results, you can reset the seed value. Monte Carlo output always displays the seed used in that analysis, so that you can reset the seed to that value if you

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.

2 Exact Tests

A fundamental problem in statistical inference is summarizing observed data in terms of a p value. The p value forms part of the theory of hypothesis testing and may be regarded an index for judging whether to accept or reject the null hypothesis. A very small p value is indicative of evidence against the null hypothesis, while a large p value implies that the observed data are compatible with the null hypothesis. There is a long tradition of using the value 0.05 as the cutoff for rejection or acceptance of the null hypothesis. While this may appear arbitrary in some contexts, its almost universal adoption for testing scientific hypotheses has the merit of limiting the number of false-positive conclusions to at most 5%. At any rate, no matter what cutoff you choose, the p value provides an important objective input for judging if the observed data are statistically significant. Therefore, it is crucial that this number be computed accurately.

Since data may be gathered under diverse, often nonverifiable, conditions, it is desirable, for p value calculations, to make as few assumptions as possible about the underlying data generation process. In particular, it is best to avoid making assumptions about the distribution, such as that the data came from a normal distribution. This goal has spawned an entire field of statistics known as nonparametric statistics. In the preface to his book, *Nonparametrics: Statistical Methods Based on Ranks*, Lehmann (1975) traces the earliest development of a nonparametric test to Arbuthnot (1710), who came up with the remarkably simple, yet popular, sign test. In this century, nonparametric methods received a major impetus from a seminal paper by Frank Wilcoxon (1945) in which he developed the now universally adopted Wilcoxon signed-rank test and the Wilcoxon rank-sum test. Other important early research in the field of nonparametric methods was carried out by Friedman (1937), Kendall (1938), Smirnov (1939), Wald and Wolfowitz (1940), Pitman (1948), Kruskal and Wallis (1952), and Chernoff and Savage (1958). One of the earliest textbooks on nonparametric statistics in the behavioral and social sciences was Siegel (1956).

The early research, and the numerous papers, monographs and textbooks that followed in its wake, dealt primarily with hypothesis tests involving continuous distributions. The data usually consisted of several independent samples of real numbers (possibly containing ties) drawn from different populations, with the objective of making distribution-free one-, two-, or K-sample comparisons, performing goodness-of-fit tests, and computing measures of association. Much earlier, Karl Pearson (1900) demonstrated that the large-sample distribution of a test statistic, based on the difference between the observed and expected counts of categorical data

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 example of a sparse 3×9 contingency table, shown in Figure 2.1, demonstrates that Fisher's concern was justified.

Figure 2.1 Sparse 3 x 9 contingency table

VAR1 * VAR2 Crosstabulation

Count

		VAR2								
		1	2	3	4	5	6	7	8	9
VAR1	1		7						1	1
	2	1	1	1	1	1	1	1		
	3		8							

The Pearson chi-square test is commonly used to test for row and column independence. For the above table, the results are shown in Figure 2.2.

Figure 2.2 Pearson chi-square test results for sparse 3 x 9 table

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	22.286[1]	16	.134

1. 25 cells (92.6%) have expected count less than 5. The minimum expected count is .29.

The observed value of the Pearson's statistic is $X^2 = 22.29$, and the asymptotic p value is the tail area to the right of 22.29 from a chi-square distribution with 16 degrees of freedom. This p value is 0.134, implying that it is reasonable to assume row and column independence. With Exact Tests, you can also compute the tail area to the right of 22.29 from the exact distribution of Pearson's statistic. The exact results are shown in Figure 2.3.

Figure 2.3 Exact results of Pearson chi-square test for sparse 9 x 3 table

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)
Pearson Chi-Square	22.286[1]	16	.134	.001

1. 25 cells (92.6%) have expected count less than 5. The minimum expected count is .29.

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 coupled with the availability of normal approximations leads to the almost automatic computation of asymptotic distributions and moments for discrete random variables. Three questions may be asked by a potential user of these asymptotic calculations:

1. How does one make them? What are the formulas and techniques for getting the answers?
2. How does one justify them? What conditions are needed to ensure that these formulas and techniques actually produce valid asymptotic results?
3. How does one relate asymptotic results to pre-asymptotic situations? How close are the answers given by an asymptotic formula to the actual cases of interest involving finite samples?

These questions differ vastly in the ease with which they may be answered. The answer to (1) usually requires mathematics at the level of elementary calculus. Question (2) is rarely answered carefully, and is typically tossed aside by a remark of the form '...assuming that higher order terms may be ignored...' Rigorous answers to question (2) require some of the deepest results in mathematical probability theory. Question (3) is the most important, the most difficult, and consequently the least answered. Analytic answers to question (3) are usually very difficult, and it is more common to see reported the result of a simulation or a few isolated numerical calculations rather than an exhaustive answer.

The concerns expressed by R. A. Fisher and by Bishop, Fienberg, and Holland can be resolved if you directly compute exact p values instead of replacing them with their asymptotic versions and hoping that these will be accurate. Fisher himself suggested the use of exact p values for 2×2 tables (1925) as well as for data from randomized experiments (1935). Exact Tests computes an exact p value for practically every important nonparametric test on either continuous or categorical data. This is achieved by permuting the observed data in all possible ways and comparing what was actually observed to what might have been observed. Thus exact p values are also known as permutational p values. The following two sections illustrate through concrete examples how the permutational p values are computed.

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.

Alternative Hypothesis: Exam results and race of examinee are not independent.

To test the hypothesis of independence, use the Pearson chi-square test of independence, available in the Crosstabs procedure. To get the results shown in Figure 2.5, the test was conducted at the 0.05 significance level:

Figure 2.5 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 the exam results are independent of the race of the examinee. However, notice that 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, X^2, is computed from the observed and the expected counts under the null hypothesis of independence as follows:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - \hat{x}_{ij})^2}{\hat{x}_{ij}} \qquad \textbf{Equation 2.1}$$

where x_{ij} is the observed count, and

$$\hat{x}_{ij} = (m_i n_j)/N \qquad \textbf{Equation 2.2}$$

is the expected count in cell (i, j) of an $r \times c$ contingency table whose row margins are $(m_1, m_2, \ldots m_r)$, column margins are $(n_1, n_2, \ldots n_c)$, and total sample size is N.

Statistical theory shows that, under the null hypothesis, the random variable X^2 *asymptotically* follows the theoretical chi-square distribution with $(r-1)\times(c-1)$ degrees of freedom. Therefore, the asymptotic p value is

$$\Pr(\chi^2 \geq 11.55556) = 0.07265 \qquad \textbf{Equation 2.3}$$

where χ^2 is a random variable following a chi-square distribution with 6 degrees of freedom.

The term **asymptotically** means "given a sufficient sample size," though it is not easy to describe the sample size needed for the chi-square distribution to approximate the exact distribution of the Pearson statistic.

One rule of thumb is:

- The minimum expected cell count for all cells should be at least 5 (Cochran, 1954). The problem with this rule is that it can be unnecessarily conservative.

Another rule of thumb is:

- For tables larger than 2×2, a minimum expected count of 1 is permissible as long as no more than about 20% of the cells have expected values below 5 (Cochran, 1954).

While these and other rules have been proposed and studied, no simple rule covers all cases. (See Agresti, 1990, for further discussion.) In our case, considering sample size, number of cells relative to sample size, and small expected counts, it appears that relying on an asymptotic result to compute a p value might be problematic.

What if, instead of relying on the distribution of χ^2, it were possible to use the true sampling distribution of X^2 and thereby produce an *exact p* value? Using Exact Tests, you can do that. The following discussion explains how this p value is computed, and why it is exact. For technical details, see Chapter 9. Consider the observed 3×4 crosstabulation (see Figure 2.4) relative to a reference set of other 3×4 tables that are like it in every possible respect, except in terms of their reasonableness under the null hypothesis. It is generally accepted that this reference set consists of all 3×4 tables of the form shown below and having the same row and column margins as Figure 2.4. (see, for example, Fisher, 1973, Yates, 1984, Little, 1989, and Agresti, 1992).

x_{11}	x_{12}	x_{13}	x_{14}	9
x_{21}	x_{22}	x_{23}	x_{24}	2
x_{31}	x_{32}	x_{33}	x_{34}	9
5	5	5	5	20

This is a reasonable choice for a reference set, even when these margins are not naturally fixed in the original data set, because they do not contain any information about the null hypothesis being tested. The exact p value is then obtained by identifying all of the tables in this reference set for which Pearson's statistic equals or exceeds 11.55556, the

observed statistic, and summing their probabilities. This is an exact *p* value because the probability of any table, $\{x_{ij}\}$, in the above reference set of tables with fixed margins can be computed exactly under the null hypothesis. It can be shown to be the hypergeometric probability

$$P(\{x_{ij}\}) = \frac{\Pi_{j=1}^{c} n_j! \Pi_{i=1}^{r} m_i!}{N! \Pi_{j=1}^{c} \Pi_{i=1}^{r} x_{ij}!} \qquad \textbf{Equation 2.4}$$

For example, the table

5	2	2	0	9
0	0	0	2	2
0	3	3	3	9
5	5	5	5	20

is a member of the reference set. Applying Equation 2.1 to this table yields a value of $X^2 = 14.67$ for Pearson's statistic. Since this value is greater than the value $X^2 = 11.55556$, this member of the reference set is regarded as more extreme than Figure 2.4. Its exact probability, calculated by Equation 2.4, is 0.000108, and will contribute to the exact *p* value. The following table

4	3	2	0	9
1	0	0	1	2
0	2	3	4	9
5	5	5	5	20

is another member of the reference set. You can easily verify that its Pearson statistic is $X^2 = 9.778$, which is less than 11.55556. Therefore, this table is regarded as less extreme than the observed table and does not count towards the *p* value. In principle, 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:

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

The exact results are shown in Figure 2.6.

Figure 2.6 Exact results of the Pearson chi-square test for fire fighter data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)
Pearson Chi-Square	11.556[1]	6	.073	.040

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

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 2×2 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.

Figure 2.7 Fisher's tea-tasting experiment

GUESS * POUR Crosstabulation

			POUR		
			Milk	Tea	Total
GUESS	Milk	Count	3	1	4
		Expected Count	2.0	2.0	4.0
	Tea	Count	1	3	4
		Expected Count	2.0	2.0	4.0
Total		Count	4	4	8
		Expected Count	4.0	4.0	8.0

Given the woman's performance in the experiment, can you conclude that she could distinguish whether milk or tea was added to the cup first? Figure 2.7 shows that she guessed correctly more times than not, but on the other hand, the total number of trials was not very large, and she might have guessed correctly by chance alone. Null and alternative hypotheses can be formulated as follows:

Null Hypothesis: The order in which milk or tea is poured into a cup and the taster's guess of the order are independent.

Alternative Hypothesis: The taster can correctly guess the order in which milk or tea is poured into a cup.

Note that the alternative hypothesis is one-sided. That is, although there are two possibilities—that the woman guesses better than average or she guesses worse than average—we are only interested in detecting the alternative that she guesses better than average.

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.

Figure 2.8 Pearson chi-square test results for tea-tasting experiment

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	2.000[2]	1	.157

2. 4 cells (100.0%) have expected count less than 5. The minimum expected count is 2.00.

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 2×2 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 2×2 table. This fact enables you to calculate an exact p value rather than rely on an asymptotic justification.

Let the generic four-fold table, $\{x_{ij}\}$, take the form

x_{11}	x_{12}	m_1
x_{21}	x_{22}	m_2
n_1	n_2	N

with $(x_{11}, x_{12}, x_{21}, x_{22})$ being the four cell counts; m_1 and m_2, the row totals; n_1 and n_2, the column totals; and N, the table total. If you assume the marginal totals as given, the value of x_{11} determines the other three cell counts. Assuming fixed marginals, the distribution of the four cell counts follows the hypergeometric distribution, stated here in terms of x_{11}:

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

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)	*p* 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
	2	2	4		
	4	4	8		
$x_{11} = 3$	3	1	4	0.229	0.243
	1	3	4		
	4	4	8		
$x_{11} = 4$	4	0	4	0.014	0.014
	0	4	4		
	4	4	8		

The probability of each possible table in the reference set of 2×2 tables with the observed margins is obtained from the hypergeometric distribution formula shown in Equation 2.6. The p values shown above are the sums of probabilities for all outcomes at least as favorable (in terms of guessing correctly) as the one in question. For example, since the table actually observed has $x_{11} = 3$, the exact p value is the sum of probabilities of all of the tables for which x_{11} equals or exceeds 3. The exact results are shown in Figure 2.9.

Figure 2.9 Exact results of the Pearson chi-square test for tea-tasting experiment

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-tailed)
Pearson Chi-Square	2.000[2]	1	.157	.486	.243

2. 4 cells (100.0%) have expected count less than 5. The minimum expected count is 2.00.

The exact result works out to $0.229 + 0.014 = 0.243$. Given such a relatively large p value, you would conclude that the woman's performance does not furnish sufficient evidence that she can correctly guess milk-tea pouring order. Note that the asymptotic p value for the Pearson chi-square test of independence was 0.079, a dramatically different number. The exact test result leads to the same conclusion as the asymptotic test result, but the exact p value is very different from 0.05, while the asymptotic p value is only marginally larger than 0.05. In this example, all 4 margins of the 2×2 table were fixed by design. For the example, in "Pearson Chi-Square Test for a 3 x 4 Table" on p. 14, the margins were not fixed. Nevertheless, for both examples, the reference set was constructed from fixed row and column margins. Whether or not the margins of the 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

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	7
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	7
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	12
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	4
x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	4
4	5	6	5	7	7	34

contains 1.6 billion tables, which presents a challenging computational problem. Fortunately, two developments have made exact *p* value computations practically feasible. First, the computer revolution has dramatically redefined what is computationally doable and affordable. Second, many new fast and efficient computational algorithms have been published over the last decade. Thus, problems that would have taken several hours or even days to solve now take only a few minutes.

It is useful to have some idea about how the algorithms in Exact Tests work. There are two basic types of algorithms: complete enumeration and Monte Carlo enumeration. The complete enumeration algorithms enumerate every single outcome in the reference set. Thus they always produce the exact *p* value. Their result is essentially 100% accurate. They are not, however, guaranteed to solve every problem. Some data sets might be too large for complete enumeration of the reference set within given time and machine limits. For this reason, Monte Carlo enumeration algorithms are also provided. These algorithms enumerate a random subset of all the possible outcomes in the reference set. The Monte Carlo algorithms provide an estimate of the exact *p* value, called the Monte Carlo *p* value, which can be made as accurate as necessary for the problem at hand. Typically, their result is 99% accurate, but you are free to increase the level of accuracy to any arbitrary degree simply by sampling more outcomes from the reference set. Also, they are guaranteed to solve any problem, no matter how large the data set. Thus, they provide a robust, reliable back-up for the situations in which the complete enumeration algorithms fail. Finally, the asymptotic *p* value is always available by default.

General guidelines for when to use the exact, Monte Carlo, or asymptotic *p* values include the following:

- It is wise to never report an asymptotic *p* value without first checking its accuracy against the corresponding exact or Monte Carlo *p* value. You cannot easily predict a priori when the asymptotic *p* value will be sufficiently accurate.
- The choice of exact versus Monte Carlo is largely one of convenience. The time required for the exact computations is less predictable than for the Monte Carlo computations. Usually, the exact computations either produce a quick answer, or else they quickly terminate with the message that the problem is too hard for the exact algorithms. Sometimes, however, the exact computations can take several hours, in which case it is better to interrupt them by selecting Stop Processor from the File menu and repeating the analysis with the Monte Carlo option. The Monte Carlo *p* values are for most practical purposes just as good as the exact *p* values. 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 other times fail to compute an exact *p* value for a data set whose sample size is less than 30. The type of exact test desired, the degree of imbalance in the allocation of subjects to treatments, the number of rows and columns in a crosstabulation, the number of ties in the data, and a variety of other factors interact in complicated ways to determine if a particular data set is amenable to exact inference. It is thus a very difficult task to specify the precise upper limits of computational feasibility for the exact algorithms. It is more useful to specify sample size and table dimension ranges within which the exact algorithms will produce quick answers—that is, *within a few seconds*. Table 1.1 and Table 1.2 describe the conditions under which exact tests can be computed quickly. In general, almost every exact test in Exact Tests can be executed in just a few seconds, provided the sample size does not exceed 30. The Kruskal-Wallis test, the runs tests, and tests on the Pearson and Spearman correlation coefficients are exceptions to this general rule. They require a smaller sample size to produce quick answers.

When to Use Monte Carlo P Values

Many data sets are too large for the exact *p* value computations, yet too sparse or unbalanced for the asymptotic results to be reliable. Figure 2.10 is an example of such a data set, taken from Senchaudhuri, Mehta, and Patel (1995). This data set reports the thickness of the left ventricular wall, measured by echocardiography, in 947 athletes participating in 25 different sports in Italy. There were 16 athletes with a wall thickness of ≥ 13mm, which is indicative of hypertrophic cardiomyopathy. The objective is to determine whether there is any correlation between presence of this condition and the type of sports activity.

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

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

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

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

Thus, a $100 \times (1-\gamma)$ % confidence interval for p is

$$CI = \hat{p} \pm z_{\gamma/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{M}} \qquad \textbf{Equation 2.9}$$

where z_α is the α th percentile of the standard normal distribution. For example, if you wanted a 99% confidence interval for p, you would use $Z_{0.005} = -2.576$. 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.

Figure 2.12 Monte Carlo results for left ventricular wall thickness versus sports activity data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Monte Carlo Significance (2-tailed)		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Likelihood Ratio	32.495	24	.115	.044[2]	.039	.050

2. Based on 10000 and seed 2000000 ...

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 Equation 2.8) and hence reduce the width of the confidence 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

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Monte Carlo Significance (2-tailed)		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Likelihood Ratio	32.495	24	.115	.045[2]	.043	.047

2. Based on 50000 and seed 2000000 ...

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, $E(\hat{p}) = p$.
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 bound of the confidence interval is above 0.05 is satisfying. Facts like these can usually be quickly established by sampling about 10,000 tables, and this takes only a few seconds.
6. The asymptotic p value carries no probabilistic guarantee whatsoever as to its accuracy. In the present example, the asymptotic p value is 0.115, implying, incorrectly, that there is no interaction between the ventricular wall thickness and the sports activity. The Monte Carlo estimate on the other hand does indeed establish this relationship at the 5% significance level.

To summarize:

- The Monte Carlo option with a sample of size 10,000 and a confidence level of 99% is the default in Exact Tests. At these default values, the Monte Carlo option provides very accurate estimates of exact p values in a just few seconds. These defaults can be easily changed in the Monte Carlo dialog box.
- Users will find that even when the width of the Monte Carlo confidence interval is wider than they'd like, the point estimate itself is very close to the exact p value. For the fire fighters data discussed in "Pearson Chi-Square Test for a 3 x 4 Table" on p. 14, the Monte Carlo estimate of the exact p value for the Pearson chi-square test is shown in Figure 2.14.

Figure 2.14 Monte Carlo results of 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 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 slightly from run to run, because you are using a different starting seed each time. However, you will also observe that each Monte Carlo estimate is very close to the exact *p* value. Thus, even if you ignored the information in the confidence interval, the Monte Carlo point estimate itself is often good enough for routine use. For a more refined analysis, you may prefer to report both the point estimate and the confidence interval.

- If you want to replicate someone else's Monte Carlo results, you need to know the starting seed used previously. Exact Tests reports the starting seed each time you run a test. If you don't specify your own starting seed, Exact Tests provides one. See "How to Set the Random Number Seed" on p. 9 in Chapter 1 for information on setting the random number seed.

When to Use Asymptotic P Values

Although the exact *p* value can be shown to converge mathematically to the corresponding asymptotic *p* value as the sample size becomes infinitely large, this property is not of much practical value in guaranteeing the accuracy of the asymptotic *p* value for any specific data set. There are many different data configurations where the asymptotic methods perform poorly. These include small data sets, data sets containing ties, large but unbalanced data sets, and sparse data sets. A numerical example follows for each of these situations.

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

subject	author	social
1	82	42
2	98	46
3	87	39
4	40	37
5	116	65
6	113	88
7	111	86

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

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

1. Based on normal approximation

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.

Figure 2.18 Two-sample Kolmogorov-Smirnov test results for diastolic blood pressure data

Frequencies

			N
Diastolic Blood Pressure	GROUP	Treated	6
		Control	7
		Total	13

Test Statistics[1]

		Diastolic Blood Pressure
Most Extreme Differences	Absolute	.667
	Positive	.667
	Negative	.000
Kolmogorov-Smirnov Z		1.198
Asymp. Sig. (2-tailed)		.113
Exact Significance (2-tailed)		.042
Point Probability		.042

1. Grouping Variable: GROUP

The asymptotic two-sided p value is 0.113. In contrast, the exact two-sided p value is 0.042, less than half the asymptotic result. The poor performance of the asymptotic test is attributable to the large number of tied observations in this data set. Suppose, for example, that the data were free of any ties, as shown in Figure 2.19.

Figure 2.19 Diastolic blood pressure of treated and control groups, without ties

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

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

Figure 2.20 Two-sample Kolmogorov-Smirnov test results for diastolic blood pressure data, without ties

Frequencies

			N
Diastolic Blood Pressure	GROUP	Treated	6
		Control	7
		Total	13

Test Statistics[1]

		Diastolic Blood Pressure
Most Extreme Differences	Absolute	.667
	Positive	.667
	Negative	.000
Kolmogorov-Smirnov Z		1.198
Asymp. Sig. (2-tailed)		.113
Exact Significance (2-tailed)		.042
Point Probability		.042

1. Grouping Variable: GROUP

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 2×5 contingency table.

Figure 2.21 Alcohol during pregnancy and birth defects

Malformation * Maternal Alcohol Consumption (drinks/day) Crosstabulation

Count

		Maternal Alcohol Consumption (drinks/day)				
		0	<1	1-2	3-5	>=6
Malformation	Absent	17066	14464	788	126	37
	Present	48	38	5	1	1

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 malformation. The results are shown in Figure 2.22.

Figure 2.22 Results of linear-by-linear association test for maternal drinking data

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Linear-by-Linear Association	1.828[2]	1	.176	.179	.105	.028

2. Standardized stat. is 1.352 ...

The asymptotic two-sided p value is 0.176. In contrast, the two-sided exact p value is 0.179.

Sparse Data Sets

Data were gathered from 250 college and university administrators on various indicators of performance like the number of applications for admittance, student/faculty ratio, faculty salaries, average SAT scores, funding available for inter-collegiate sports, and so forth. Figure 2.23 shows a crosstabulation of competitiveness against the student/faculty ratio for a subset consisting of the 65 state universities that participated in the survey.

Figure 2.23 Student/faculty ratio versus competitiveness of state universities

Student/Faculty Ratio * Competitiveness of Institution Crosstabulation

Count

		Competitiveness of Institution					
		Less	Average	Very	Highly	Most	Total
Student/Faculty Ratio	2				1		1
	7		1		1		2
	8		1			1	2
	9		1				1
	10	1		2			3
	11	1	3		1		5
	12		2	1			3
	13	1	3	1			5
	14	3	3	1			7
	15	1	5	1	1		8
	16	1	5				6
	17	3	2	1			6
	18		2	4	1		7
	20		2				2
	21	2					2
	22			1			1
	23		1				1
	24		1	1			2
	70		1				1
Total		13	33	13	5	1	65

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

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	94.424[1]	72	.039	.114[2]	.106	.122

1. 95 cells (100.0%) have expected count less than 5. The minimum expected count is .02.
2. Based on 10000 and seed 2000000 ...

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, however, that the table, though large, is very sparse. Because this data set is so large, the Monte Carlo result, rather than the exact result, is shown. The Monte Carlo estimate of the exact *p* value is 0.114. This is a three-fold increase in the *p* value, which suggests that there is, after all, no interaction between competitiveness and the student/faculty ratio at state universities.

It should be clear from the above examples that it is very difficult to predict a priori if a given data set is large enough to rely on an asymptotic approximation to the *p* value. The notion of what constitutes a large sample depends on the structure of the data and the test being used. It cannot be characterized by any single measure. A crosstabulation created from several thousand observations might nevertheless produce inaccurate asymptotic *p* values if it possesses many cells with small counts. On the other hand, a rank test like the Wilcoxon, performed on continuous, well-balanced data, with no ties, could produce an accurate asymptotic *p* value with a sample size as low as 20. Ultimately, the best definition of a large data set is an operational one—if a data set produces an accurate asymptotic *p* value, it is large; otherwise, it is small. In the past, such a definition would have been meaningless, since there was no gold standard by which to gauge the accuracy of the asymptotic *p* value. In Exact Tests, however, either the exact *p* value or its Monte Carlo estimate is readily available to make the comparison and may be used routinely along with the asymptotic *p* value.

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 $1 \times c$ contingency table, where c is the number of categories. Cell i of this $1 \times c$ table contains a frequency count, O_i, of the number of observations falling into category i. Along the bottom of the table is a $(1 \times c)$ vector of cell probabilities

$$\pi = (\pi_1, \pi_2, \ldots \pi_c) \qquad \textbf{Equation 3.1}$$

such that π_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**	**...**	***c***	
Cell Counts	O_1	O_2	...	O_c	N
Cell Probabilities	π_1	π_2	...	π_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 π. The null hypothesis

$$H_0{:}(O_1, O_2, \ldots O_c) \sim \text{Multinomial}(\pi, N) \qquad \textbf{Equation 3.2}$$

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

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

where $E_i = N\pi_i$ 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 $N \to \infty$, 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

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

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 π_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

$$p_2 = \Pr(\chi^2 \geq x^2)$$ **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

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

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

$$CI = \hat{p}_2 \pm 2.576\sqrt{(\hat{p}_2)(1-\hat{p}_2)/M}$$ **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 σ is based on inverting an exact binomial hypothesis test when an extreme outcome is encountered. If $\hat{p}_2 = 0$, an $\alpha\%$ confidence interval for the exact *p* value is

$$CI = [0, 1-(1-\alpha/100)^{1/M}]$$ **Equation 3.8**

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

$$CI = [(1-\alpha/100)^{1/M}, 1]$$ **Equation 3.9**

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.

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

$$\tilde{p}_2 = (\Pr\chi_3^2 \geq 8.0) = 0.046$$

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

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

Exact Tests also provides the point probability that the test statistic equals 8.0. This probability, 0.0203, is a measure of the discreteness of the exact distribution of χ^2. Some statisticians advocate subtracting half of this point probability from the exact *p* value, and call the result the mid-*p* value.

Because of its small size, this data set does not require a Monte Carlo analysis. However, results obtained from a Monte Carlo analysis are more accurate than results produced by an asymptotic analysis. Figure 3.2 shows the Monte Carlo estimate of the exact *p* value based on a Monte Carlo sample of 10,000.

Figure 3.2 Monte Carlo results for chi-square test

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.	Monte Carlo Sig.		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
CATEGORY	8.000	3	.046	.049[2]	.044	.055

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

2. Based on 10000 sampled tables with starting seed 2000000.

The Monte Carlo estimate of the exact p value is 0.0493, which is much closer to the exact *p* value of 0.0523 than the asymptotic result. But a more important benefit of the Monte Carlo analysis is that we also obtain a 99% confidence interval. In this example, with a Monte Carlo sample of 10,000, the interval is (0.0437, 0.0549). This interval could be narrowed by sampling more multinomial vectors from the null distribution. To obtain more conclusive evidence that the exact *p* value exceeds 0.05 and thus is not statistically

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

Figure 3.3 Monte Carlo results for chi-square test with 100,000 samples

Test Statistics

	Chi-Square[1]	df	Asymp. Sig.	Monte Carlo Sig.		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
CATEGORY	8.000	3	.046	.051[2]	.049	.053

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

2. Based on 100000 sampled tables with starting seed 2000000.

This time, the Monte Carlo estimate is 0.0508, almost indistinguishable from the exact result. Moreover, the exact *p* value is guaranteed, with 99% confidence, to lie within the interval (0.0490, 0.0525). We are now 99% certain that the exact *p* value exceeds 0.05.

Example: A Medium-Sized Data Set

This example shows that the chi-square approximation may not be reliable even when the sample size is as large as 50, has only three categories, and has cell counts that satisfy the Siegel and Castellan criteria discussed on p. 40. Table 3.4 displays data from Radlow and Alt (1975) showing observed counts and multinomial probabilities under the null hypothesis for a multinomial distribution with three categories.

Table 3.4 Frequency counts from a multinomial distribution with three categories

	Multinomial Categories			Row Total
	1	**2**	**3**	
Cell counts	12	7	31	50
Cell Probabilities	0.2	0.3	0.5	1

Figure 3.4 shows the results of the chi-square goodness-of-fit test on these data.

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.

The data consist of N i.i.d. observations, $(u_i,u_2, \ldots u_N)$, from an unknown distribution $G(u)$; i.e. $G(u) = \Pr(U \leq u)$. Let $F(u)$ be a completely specified distribution. The Kolmogorov test is used to test the null hypothesis

$$H_0: G(u) = F(u) \text{ for all } u \qquad \textbf{Equation 3.10}$$

H_0 can be tested against either a two-sided alternative or a one-sided alternative. The two-sided alternative is

$$H_1: G(u) \neq F(u) \text{ for at least one value of } u \qquad \textbf{Equation 3.11}$$

Two one-sided alternative hypotheses can be specified. One states that *F* is stochastically greater than *G*. That is,

$$H_{1a}: G(u) < F(u) \text{ for at least one value of } u \qquad \textbf{Equation 3.12}$$

The other one-sided alternative states the complement, that *G* is stochastically greater than *F*. That is,

$$H_{1b}: F(u) < G(u) \text{ for at least one value of } u \qquad \textbf{Equation 3.13}$$

The test statistics for testing H_0 against either H_1, H_{1a}, or H_{1b} are all functions of the specified distribution, $F(u)$, and the empirical cumulative density function (c.d.f.), $S(u)$, is derived from the observed values, $(u_i,u_2, \ldots u_N)$. The test statistic for testing H_0 against H_1 is

$$T = \sup_u \{|F(u) - S(u)|\} \qquad \textbf{Equation 3.14}$$

The test statistic for testing H_0 against H_{1a} is

$$T^+ = \sup_u \{F(u) - S(u)\} \qquad \textbf{Equation 3.15}$$

The test statistic for testing H_0 against H_{1b} is

$$T^- = \sup_u \{S(u) - F(u)\} \qquad \textbf{Equation 3.16}$$

Kolmogorov derived asymptotic distributions as $N \to \infty$, for *T*, T^+, and T^-. For small *N*, the exact *p* values provided by Exact Tests are appropriate. If $F(u)$ is a discrete distribution, the exact *p* values can be computed using the method described by Conover (1980). If $F(u)$ is a continuous distribution, the exact *p* value can be computed using the results given by Durbin (1973).

Example: Testing for a Uniform Distribution

This example is taken from Conover (1980). A random sample size of 10 is drawn from a continuous distribution. The sample can be tested to determine if it came from a uniform continuous distribution with limits of 0 and 1. Figure 3.5 shows the data displayed n the Data Editor.

Figure 3.5 Data to test for a uniform distribution

observ	value
1	.621
2	.503
3	.203
4	.477
5	.711
6	.581
7	.329
8	.480
9	.554
10	.382

We can run the Kolmogorov-Smirnov test to determine if the sample was generated by a uniform distribution. The results are displayed in Figure 3.6.

Figure 3.6 Kolmogorov-Smirnov results

One-Sample Kolmogorov-Smirnov Test

	N	Uniform Parameters[1,2]		Most Extreme Differences			Kolmogorov-Smirnov Z	Asymp. Sig. (2-tailed)	Exact Significance (2-tailed)	Point Probability
		Minimum	Maximum	Absolute	Positive	Negative				
VALUE	10	0	0	.289	.289	-.229	.914	.374	.311	.000

1. Test distribution is Uniform.

2. User-Specified

The exact exact two-sided *p* value is 0.311. The asymptotic two-sided *p* value is 0.3738.

4 One-Sample Inference for Binary Data

This chapter discusses two statistical procedures for analyzing binary data in Exact Tests. First, it describes exact hypothesis testing and exact confidence interval estimation for a binomial probability. Next, it describes the runs test (also known as the Wald-Wolfowitz one-sample runs test) for determining if a sequence of binary observations is random. You will see that although the theory underlying the runs test is based on a binary sequence, the test itself is applied more generally to non-binary observations. For this reason, the data are transformed automatically in Exact Tests from a non-binary to a binary sequence prior to executing the test.

Available Tests

Table 4.1 shows the tests for binary data available in Exact Tests, the procedure from which each can be obtained, and a bibliographical reference for each.

Table 4.1 Available tests

Test	Procedure	Reference
Binomial test	Nonparametric Tests: Binomial Test	Conover (1971)
Runs test	Nonparametric Tests: Runs Test	Lehmann (1975

Binomial Test and Confidence Interval

The data consist of t successes and $N-t$ failures in N independent Bernoulli trials. Let π be the true underlying success rate. Then the outcome $T = t$ has the binomial probability

$$\Pr(T = t|\pi) = \binom{N}{t}\pi^t(1-\pi)^{N-t} \qquad \textbf{Equation 4.1}$$

Exact Tests computes the observed proportion $\hat{\pi}$, which is also the maximum-likelihood estimate of π, as

$$\hat{\pi} = t/N$$

To test the null hypothesis

$$H_0: \pi = \pi_o \qquad \textbf{Equation 4.2}$$

Exact Tests computes the following one- and two-sided p values:

$$p_1 = \min\{\Pr(T \le t|\pi_o), \Pr(T \ge t|\pi_o)\} \qquad \textbf{Equation 4.3}$$

and

$$p_2 = 2*p_1 \qquad \textbf{Equation 4.4}$$

Example: Pilot Study for a New Drug

Twenty patients were treated in a pilot study of a new drug. There were four responders (successes) and 16 non-responsive (failures). The binomial test can be run to test the null hypothesis that $\pi = 0.05$.

These data can be entered into the Data Editor using a response variable with 20 cases. If *successes* are coded as 1's, and *failures* are coded as 0's, *response* contains sixteen cases with a value of 0, and four cases with a value of 1.

The binomial test performed on these data produces the results displayed in Figure 4.1.

Figure 4.1 Binomial test results for drug study

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)	Point Probability
Response to Drug	Group 1	Success	4	.2	.05	.016	.013
	Group 2	Failure	16	.80			
	Total		20	1.00			

The exact one-sided p value is 0.0159, so the null hypothesis that $\pi = 0.05$ is rejected at the 5% significance level.

Runs Test

Consider a sequence of N binary outcomes, $(y_i, y_2, \ldots y_N)$, where each y_i 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 of one 0. Finally, there is a run of one 1. In all, there are seven runs in the above sequence. Let the random variable R denote the number of runs in a binary sequence consisting of m 1's and n 0's, where $m + n = N$. The Wald-Wolfowitz runs test is used to test the null hypothesis

H_0: The sequence of m 1's and n 0's, $(m + n) = N$, was generated by N independent Bernoulli trials, each with a probability π of generating a 1 and a probability $(1 - \pi)$ of generating a 0.

Very large or very small values of R are evidence against H_0. In order to determine what constitutes a very large or a very small run, the distribution of R is needed. Although unconditionally the distribution of R depends on π, this nuisance parameter can be eliminated by working with the conditional distribution of R, given that there are m 1's and n 0's in the sequence. This conditional distribution can be shown to be

$$\Pr(R = 2k) = \frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{N}{n}} \qquad \textbf{Equation 4.5}$$

and

$$\Pr(R = 2k+1) = \frac{\binom{m-1}{k-1}\binom{n-1}{k} + \binom{m-1}{k}\binom{n-1}{k-1}}{\binom{N}{n}} \qquad \textbf{Equation 4.6}$$

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

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

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

$$z = \frac{r + h - (2mn/N) - 1}{\sqrt{[2mn(2mn - N)]/[N^2(n-1)]}} \qquad \textbf{Equation 4.8}$$

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 $x_1, x_2, \ldots x_N$, where the x_i's are not binary. In that case, the approach suggested by Lehmann (1975) is to replace each x_i with a corresponding binary transformation

$$y_i = \begin{cases} 1 \text{ if } y_i \geq \tilde{x} \\ 0 \text{ if } y_i < \tilde{x} \end{cases} \qquad \textbf{Equation 4.9}$$

where $\tilde{x}$ is the median of the observed data series. The median is calculated in the following way. Let $x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[N]}$ be the observed data series sorted in ascending order. Then

$$\tilde{x} = \begin{matrix} x_{[(N+1)/2]} \text{ if } N \text{ is odd} \\ (x_{[N/2]} + x_{[(N+2)/2]})/2 \text{ if } N \text{ is even} \end{matrix} \qquad \textbf{Equation 4.10}$$

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

probability of obtaining 10 or fewer runs, is 0.301 and does not indicate any significant departure from randomness.

If the data set had been larger, it would have been difficult to compute the exact test, and you would have had to either rely on the asymptotic results or estimate the exact *p* values using the Monte Carlo option. Figure 4.4 shows Monte Carlo estimates of the exact *p* values for the runs test based on 10,000 random permutations of the 12 0's and 12 1's in a binary sequence of 24 numbers. Each permutation is assigned an equally likely probability given by $24!/(12!12!) = (1/2704156)$.

Figure 4.4 Monte Carlo results for runs test for aggression scores data

	Test Value[1]	Cases < Test Value	Cases >= Test Value	Total Cases	Number of Runs	Z	Asymp. Sig. (2-tailed)	Monte Carlo Sig. (2-tailed)		
								Sig.	99% Confidence Interval	
									Lower Bound	Upper Bound
SCORE	25.00	12	12	24	10	-1.044	.297	.298[2]	.286	.310

1. Median
2. Based on 10000 sampled tables with starting seed 200000.

Notice that the Monte Carlo two-sided *p* value, 0.298, is extremely close to the exact *p* value, 0.310. But more importantly, the Monte Carlo method produces a 99% confidence interval within which the exact two-sided *p* value is guaranteed to lie. In this example, the interval is (0.286, 0.310), which again demonstrates conclusively that the null hypothesis of a random data series cannot be rejected.

Example: Small Data Set

Here is a small hypothetical data set illustrating the difference between the exact and asymptotic inference for the runs test. The data consists of a binary sequence of ten observations

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

with six 1's and four 0's. Thus, there are 3 runs in this sequence. The results of the runs test are displayed in Figure 4.5.

Figure 4.5 Runs test results for small data set

	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

1. Median

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

5 Two-Sample Inference: Paired Samples

The tests in this section are commonly applied to matched pairs of data, such as when several individuals are being studied and two repeated measurements are taken on each individual. The objective is to test the null hypothesis that both measurements came from the same population. The inference is complicated by the fact that the two observations on the same individual are correlated, while there is independence across the different individuals being studied. In this setting, Exact Tests provides statistical procedures for both continuous and categorical data. For matched pairs of continuous data (possibly with ties) Exact Tests provides the sign test and the Wilcoxon signed-ranks test. For matched pairs of binary outcomes, Exact Tests provides the McNemar test. For matched pairs of ordered categorical outcomes, Exact Tests generalizes from the McNemar test to the marginal homogeneity test.

Available Tests

Table 5.1 shows the available tests for paired samples, the procedure from which they can be obtained, and a bibliographical reference for each test.

Table 5.1 Available tests

Test	Procedure	Reference
Sign test	Nonparametric Tests: Two-Related-Samples Tests	Sprent (1993)
Wilcoxon signed-ranks test	Nonparametric Tests: Two-Related-Samples Tests	Sprent (1993)
McNemar test	Nonparametric Tests: Two-Related-Samples Tests	Siegel and Castellan (1988)
Marginal homogeneity test	Nonparametric Tests: Two-Related-Samples Tests	Agresti (1990)

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. This test 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 measures** 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).

Marginal homogeneity test. This test generalizes the McNemar test from binary response to multinomial response. Specifically, it tests the equality of two $c \times 1$ multinomial response vectors. Technically, the response could be ordered or unordered. However, the methods developed in the present release of Exact Tests apply only to ordered response. The data consist of paired, dependent responses, one from population 1 and the other from population 2. Each response falls into one of c ordered categories. The data are arranged in the form of a square $c \times c$ contingency table in which an entry in cell (i, j) signifies that the response of one member of the dependent pair fell into category i, while the response of the second member fell into category j. A typical application of the test of marginal homogeneity is a repeated measures situation in which each subject's ordered categorical response is elicited twice, once before and once after a specified event (treatment) occurs. The test then determines if the response rates in the c ordered categories are altered by the treatment. See Agresti (1990) for various model-based approaches to this problem. Exact Tests provides a nonparametric solution using the generalized Mantel-Haenszel approach suggested by Kuritz, Landis, and Koch (1988). See also White, Landis, and Cooper (1982).

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 2^N 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, where the X and Y random variables are correlated, usually through a matched-pairs design. Define the N differences

$$d_i = x_i - y_i,\ i = 1, 2, \ldots, N$$

Omit from further consideration all pairs with a zero difference. Assume that for all $i, |d_i| > 0$. The following assumptions are made about the distribution of the random variables D_i:

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

Let the common median of the N D_i's be denoted by λ. The null hypothesis is

$$H_0: \lambda = 0$$

There are two one-sided alternative hypotheses of the form

$$H_1: \lambda > 0$$

and

$$H'_1: \lambda < 0$$

The two-sided alternative hypothesis is that either H_1 or H'_1 holds, but you cannot specify which.

To test these hypotheses, utilize permutational distributions of tests statistics derived from either the signs or the signed ranks of the N differences. Let the absolute values of the observed paired differences, arranged in ascending order, be

$$|d_{[1]}| \le |d_{[2]}| \ldots \le |d_{[N]}|$$

and let

$$r_{[1]} \le r_{[2]} \ldots \le r_{[N]}$$

be corresponding ranks (mid-ranks in the case of tied data). Specifically, if there are m_j observations tied at the jth smallest absolute value, you assign to all of them the rank

$$r_{[j]} = m_1 + \ldots + m_{j-1} + 1/2(m_j + 1) \qquad \textbf{Equation 5.1}$$

For the Wilcoxon signed-ranks test, inference is based on the permutational distribution of the test statistic

$$T_{SR} = \min\left\{ \sum_{i=1}^{N} r_i I(D_i > 0), \sum_{i=1}^{N} r_i I(D_i < 0) \right\} \qquad \textbf{Equation 5.2}$$

whose observed value is

$$t_{SR} = \min\left\{ \sum_{i=1}^{N} r_i I(d_i > 0), \sum_{i=1}^{N} r_i I(D_i < 0) \right\} \qquad \textbf{Equation 5.3}$$

where $I(\cdot)$ is the indicator function. It assumes a value of 1 if its argument is true and 0 otherwise. In other words, t_{SR} is the minimum of ranks of the positive differences and the ranks of the negative differences among the N observed differences.

Sometimes you do not know the actual magnitude of the difference but only have its sign available to us. In that case, you cannot rank the differences and so compute the Wilcoxon signed-ranks statistic. However, you can still use the information present in the sign of the difference and perform the sign test. For the sign test, inference is based on the permutational distribution of the test statistic

$$T_S = \min\left\{ \sum_{i=1}^{N} I(D_i > 0), \sum_{i=1}^{N} r_i I(D_i < 0) \right\} \qquad \textbf{Equation 5.4}$$

whose observed value is

$$t_s = \min\left\{\sum_{i=1}^{N} I(d_i > 0), \sum_{i=1}^{N} r_i I(D_i < 0)\right\} \quad \textbf{Equation 5.5}$$

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

The permutational distributions of T_{SR} and T_S under the null hypothesis are obtained by assigning positive or negative signs to the N differences in all possible ways. There are 2^N such possible assignments, corresponding to the reference set

$$\Gamma = \{(\text{sgn}(D_1), \text{sgn}(D_2), \ldots \text{sgn}(D_N)) : \text{sgn}(D_i) = 1 \text{ or } -1, \text{for} i = 1, 2, \ldots N\} \quad \textbf{Equation 5.6}$$

and each assignment has equal probability, 2^{-N}, under the null hypothesis. Exact Tests uses network algorithms to enumerate the reference set in Equation 5.6 in order to compute exact p values.

From Equation 5.2 and standard binomial theory, the mean of T_{SR} is

$$E(T_{SR}) = \sum_{i=1}^{N} r_i/2 \quad \textbf{Equation 5.7}$$

and the variance of T_{SR} is

$$\sigma^2(T_{SR}) = \sum_{i=1}^{N} r_i^2/4 \quad \textbf{Equation 5.8}$$

From Equation 5.4 and standard binomial theory, the mean of T_S is

$$E(T_S) = N/2 \quad \textbf{Equation 5.9}$$

and the variance of T_S is

$$\sigma^2(T_S) = N/4 \quad \textbf{Equation 5.10}$$

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 T_{SR} when the Wilcoxon signed-ranks test is being computed and denotes T_S when the sign test is being computed. In either case, you can now denote the standardized test statistic as

$$Z = \frac{T - E(t)}{\sigma(T)} \qquad \textbf{Equation 5.11}$$

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

$$\tilde{p}_2 = 2\tilde{p}_1 \qquad \textbf{Equation 5.12}$$

The exact one-sided p value is defined as

$$p_1 = \begin{cases} \Pr(T \geq t) \text{ if } t > E(T) \\ \Pr(T \leq t) \text{ if } t \leq E(T) \end{cases} \qquad \textbf{Equation 5.13}$$

where t is the observed value of T. The potential to misinterpret a one-sided p value applies in the exact setting, as well as in the asymptotic case. The exact two-sided p value is defined to be double the exact one-sided p value:

$$p_2 = 2p_i \qquad \textbf{Equation 5.14}$$

This is a reasonable definition, since the exact permutational distribution of T is symmetric about its mean.

The one-sided Monte-Carlo p value is obtained as follows. First, suppose that $t > E(T)$, so that you are estimating the right tail of the exact distribution. You sample M times from the reference set (Γ) of 2^N possible assignments of signs to the ranked data. Suppose that the ith sample generates a value t_i for the test statistic. Define the random variable

$$Z_i = \begin{cases} 1 \text{if } t_i \geq t \\ 0 \text{ otherwise} \end{cases}$$

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

$$\hat{p}_1 = \sum_{i=1}^{M} Z_i / M \qquad \textbf{Equation 5.15}$$

Next, if $t < E(T)$, so that you are estimating the left tail of exact distribution, the random variable is defined by

$$Z_i = \begin{cases} 1 \text{if } (t_i \le t) \\ 0 \text{ otherwise} \end{cases}$$

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

$$CI = \hat{p}_1 \pm 2.576\sqrt{(\hat{p}_1)(1-\hat{p}_1)/M} \qquad \textbf{Equation 5.16}$$

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 $\hat{p} = 0$ or $\hat{p} = 1$. Now the sample standard deviation is 0, but the data do not support a confidence interval of zero width. An alternative approach in this extreme situation is to invert an exact binomial hypothesis test. It can be easily shown that if $\hat{p} = 0$, an α% confidence interval for the exact p value is

$$CI = [0, 1 - (1 - \alpha/100)^{1/M}] \qquad \textbf{Equation 5.17}$$

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

$$CI = [(1 - \alpha/100)^{1/M}, 1] \qquad \textbf{Equation 5.18}$$

By symmetry, the two-sided Monte Carlo p value is twice the one-sided p value:

$$\hat{p}_2 = 2\hat{p}_1 \qquad \textbf{Equation 5.19}$$

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-sided Monte Carlo p value. The confidence interval for the true two-sided p value can thus be adjusted appropriately, based on the increased variance.

Example: AZT for AIDS

The data shown in Figure 5.1, from Makutch and Parks (1988), document the response of serum antigen level to AZT in 20 AIDS patients. Two sets of antigen levels are provided for each patient: pre-treatment, represented by *preazt*, and post-treatment, represented by *postazt*.

Figure 5.1 Response of serum antigen level to AZT

id	preazt	postazt
1	149	0
2	0	51
3	0	0
4	259	385
5	106	0
6	255	235
7	0	0
8	52	0
9	340	48
10	0	0
11	180	77
12	0	0
13	84	0
14	89	0
15	212	53
16	554	150
17	500	0
18	424	165
19	112	98
20	2600	0

Figure 5.2 shows the results for the Wilcoxon signed-ranks test.

Figure 5.2 Wilcoxon signed-ranks test results for AZT data

Ranks

		N	Mean Rank	Sum of Ranks
Serum Antigen Level Post AZT - Serum Antigen Level (pg/ml) Pre-AZT	Negative Ranks	2[1]	6.00	12.00
	Positive Ranks	14[2]	8.86	124.00
	Ties	4[3]		
	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

Test Statistics[1]

	Z	Asymp. Sig. (2-tailed)	Exact Significance (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Serum Antigen Level Post AZT - Serum Antigen Level (pg/ml) Pre-AZT	-2.896[2]	.004	.002	.001	.000

1. Wilcoxon Signed Ranks Test
2. Based on negative ranks.

The test statistic is the smaller of the two sums of ranks, which is 12. The exact one-sided p value is 0.001, about half the size of the asymptotic one-sided p value. To obtain the asymptotic one-sided p value, divide the asymptotic two-sided p value, 0.004, by 2 ($(0.004)/2 = 0.002$). If this data set had been extremely large, you might have preferred to compute the Monte Carlo estimate of the exact p value. The Monte Carlo estimate shown in Figure 5.3 is based on sampling 10,000 times from the reference set Γ, defined by Equation 5.6.

Figure 5.3 Monte Carlo results of Wilcoxon signed-ranks test for AZT data

Test Statistics[1,2]

	Z	Asymp. Sig. (2-tailed)	Monte Carlo Sig. (2-tailed)			Monte Carlo Sig. (1-tailed)		
			Sig.	99% Confidence Interval		Sig.	99% Confidence Interval	
				Lower Bound	Upper Bound		Lower Bound	Upper Bound
Serum Antigen Level Post AZT - Serum Antigen Level (pg/ml) Pre-AZT	-2.896[3]	.004	.002	.001	.004	.001	.0002	.0018

1. Wilcoxon Signed Ranks Test
2. Based on 10000 sampled tables with starting seed 2000000.
3. Based on negative ranks.

The Monte Carlo point estimate of the exact one-sided p value is 0.001, very close to the exact answer. Also, the Monte Carlo confidence interval guarantees with 99% confidence that the true p value is in the range (0.0002, 0.0018). This guarantee is unavailable with the asymptotic method; thus, the Monte Carlo estimate would be the preferred option for large samples.

Next, the exact sign test is run on these data. The results are displayed in Figure 5.4.

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

Test Statistics[1]

		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

1. Sign Test
2. Exact results are provided instead of Monte Carlo for this test.
3. Binomial distribution used.

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

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

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

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

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

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

Under the null hypothesis, the conditional probability, $P(y)$, of observing any $y \in \Gamma$ is binomial with parameters $(0.5, N_d)$. Thus,

$$P(y) = \frac{(0.5)^{N_d} N_d!}{y_{12}! y_{21}!} \qquad \textbf{Equation 5.22}$$

and the probability that the McNemar statistic equals or exceeds its observed value $MC(x)$, is readily evaluated as

$$\Pr(MC(y) \geq MC(x)) = \sum_{MC(y) \geq MC(x)} P(y) \qquad \textbf{Equation 5.23}$$

the sum being taken over all $y \in \Gamma$. The probability that the McNemar statistic is less than or equal to $MC(x)$ is similarly obtained. The exact one-sided p value is then defined as

$$p_1 = \min\{\Pr(MC(y) \leq MC(x)), \Pr(MC(y) \geq MC(x))\} \qquad \textbf{Equation 5.24}$$

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

$$p_2 = 2p_1 \qquad \textbf{Equation 5.25}$$

In large samples, the two-sided asymptotic p value is calculated by a χ^2 approximation with a continuity correction, and 1 degree of freedom, as shown in Equation 5.26.

$$\chi^2 = \frac{(|y_{12} - y_{21}| - 1)^2}{N_d} \qquad \textbf{Equation 5.26}$$

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.

$$MC(y) = y_{12} - y_{21} \qquad \textbf{Equation 5.27}$$

You should examine the one-sided p value as well as the sign of the test statistic before drawing conclusions from the data.

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
2. Binomial distribution used.

The exact one-sided p value is 0.132. Notice that the value of the McNemar statistic, $13 - 7$, has a positive sign. This indicates that of the 20 ($13 + 7$) 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 $MC(y) = MC(x) = 13 - 7 = 6$.

Marginal Homogeneity Test

The marginal homogeneity test (Agresti, 1990) is an extension of the McNemar test from two categories to more than two categories. The data are thus defined on a square $c \times c$ contingency table in which the row categories represent the first member of a pair of correlated observations, and the column categories represent the second member of the pair. In Exact Tests, the categories are required to be ordered. The data are thus represented by a $c \times c$ contingency table with entry (x_{ij}) in row i and column j. This entry is the count of the number of pairs of observations in which the first member of the pair falls into ordered category i and the second member into ordered category j. Let π_j be the probability that the first member of the matched pair falls in row j. Let π'_j be the probability that the second member of the matched pair falls in column j. The null hypothesis of marginal homogeneity states that

$$H_0: \pi_j = \pi'_j, \text{ for all } j = 1, 2, \ldots c$$

In other words, the probability of being classified into category j is the same for the first as well as the second member of the matched pair.

The marginal homogeneity test for ordered categories can be formulated as a stratified $2 \times c$ contingency table. The theory underlying this test, the definition of its test statistic, and the computation of one- and two-sided p values are discussed in Kuritz, Landis, and Koch (1988).

Example: Matched Case-Control Study of Endometrial Cancer

Figure 5.7, taken from the Los Angeles Endometrial Study (Breslow and Day, 1980), displays a crosstabulation of average doses of conjugated estrogen between cases and matched controls.

Figure 5.7 Crosstabulation of dose for cases with dose for controls

Dose (Cases) * Dose (Controls) Crosstabulation

Count

		Dose (Controls)			
		.0000	.2000	.5125	.7000
Dose (Cases)	.0000	6	2	3	1
	.2000	9	4	2	1
	.5125	9	2	3	1
	.7000	12	1	2	1

In this matched pairs setting, the test of whether the cases and controls have the same exposure to estrogen, is equivalent to testing the null hypothesis that the row margins and column margins come from the same distribution. The results of running the exact marginal homogeneity test on these data are shown in Figure 5.8.

Figure 5.8 Marginal homogeneity results for cancer data

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
Dose (Cases) & Dose (Controls)	4	45	6.687	12.869	1.655	-3.735	.000	.000	.000	.000

The *p* values are extremely small, showing that the cases and controls have significantly different exposures to estrogen. The null hypothesis of marginal homogeneity is rejected.

Example: Pap-Smear Classification by Two Pathologists

This example is taken from Agresti (1990). Two pathologists classified the Pap-smear slides of 118 women in terms of severity of lesion in the uterine cervix. The classifications fell into five ordered categories. *Level 1* is negative, *Level 2* is atypical squamous hyperplasia, *Level 3* is carcinoma in situ, *Level 4* is squamous carcinoma, and *Level 5* is invasive carcinoma. Figure 5.9 shows a crosstabulation of level classifications between two pathologists.

Figure 5.9 Crosstabulation of Pap-smear classifications by two pathologists

First Pathologist * Pathologist 2 Crosstabulation

Count

		Pathologist 2				
		Level 1	Level 2	Level 3	Level 4	Level 5
First Pathologist	Level 1	22	2	2		
	Level 2	5	7	14		
	Level 3		2	36		
	Level 4		1	14	7	
	Level 5			3		3

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.

6 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 make any assumptions about the underlying populations from which these subjects were drawn. Lehmann (1975) has demonstrated clearly that the same statistical methods are applicable whether the data arose from a population model or a randomization model. Thus, no distinction will be made between the two ways of gathering the data.

There are important differences between the structure of the data for this chapter and the previous one. The data in this chapter are independent both within a sample and across the two samples, whereas the data in the previous chapter consisted of N matched (correlated) pairs of observations with independence across pairs. Moreover, in the previous chapter, the sample size was required to be the same for each sample, whereas in this chapter, the sample size may differ, with n_j being the size of sample $j = 1, 2$.

Available Tests

Table 6.1 shows the available tests for two independent samples, the procedure from which they can be obtained, and a bibliographical reference for each test.

Table 6.1 Available tests

Test	Procedure	Reference
Mann-Whitney test	Nonparametric Tests: Two Independent Samples	Sprent (1993)
Kolmogorov-Smirnov test	Nonparametric Tests: Two Independent Samples	Conover (1980)
Wald-Wolfowitz runs test	Nonparametric Tests: Two Independent Samples	Gibbons (1985)

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 not be as powerful as the Mann-Whitney test if the alternative is that one distribution is shifted with respect to the other. One-sided forms of the Kolmogorov-Smirnov test can be specified and are powerful against the one-sided alternative that one distribution is stochastically larger than the other.

Wald-Wolfowitz runs test. The Wald-Wolfowitz runs test is a competitor to the Kolmogorov-Smirnov test for testing the equality of two distributions against general alternatives. It will not be powerful against specific alternatives such as the shift alternative, but it is a good test when no particular alternative hypothesis can be specified. This test is even more general than the Kolmogorov-Smirnov test in the sense that it has no one-sided version.

Statistical Methods

The data for all of the tests in this chapter consist of two independent samples, each of size n_j, $j = 1, 2$, where $n_1 + n_2 = N$. These N observations can be represented in the form of the one-way layout shown in Table 6.2.

Table 6.2 One-way layout for two independent samples

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

This table, denoted by *u*, displays the observed one-way layout of raw data. The observations in *u* arise from continuous univariate distributions (possibly with ties). Let the formula

$$F_j(v) = \Pr(V \le v | j), j = 1, 2 \qquad \textbf{Equation 6.1}$$

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

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

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_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 they occupy different rows and columns. In order to develop this idea more precisely, let the set *W* denote the collection of all possible 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

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

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. 80, 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.

Table 6.6 Permutation of the observed one-way layout of scores

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

All of the test statistics in this chapter are univariate functions of $\tilde{w} \in W$. Let the test statistic be denoted by $T(\tilde{w}) \equiv T$ and its observed value be denoted by $t(w) \equiv t$. The functional form of $T(\tilde{w})$ will be defined separately for each test, in subsequent sections of this chapter. Following is a discussion of how the null distribution of T may be derived in general, and how it is used for p value computations.

The Null Distribution of T

In order to test the null hypothesis, H_0, you need to derive the distribution of T under the assumption that H_0 is true. This distribution is obtained by the following permutational argument:

If H_0 is true, every member $\tilde{w} \in W$ 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 two populations or by assigning N subjects to two treatments in accordance with a predetermined randomization rule. No distinction is made between these two ways of gathering the data, although one usually applies to observational studies and the other to randomized clinical trials.

It follows from the above permutational argument that the exact probability of observing any $\tilde{w} \in W$ is

$$h(\tilde{w}) = \frac{\Pi_{i=1}^{2} n_1!}{N!} \qquad \textbf{Equation 6.4}$$

which does not depend on the specific way in which the original one-way layout, w, was permuted. Then

$$\Pr(T = t) = \sum_{T(\tilde{w}) = t} h(\tilde{w}) \qquad \textbf{Equation 6.5}$$

the sum being taken over all $\tilde{w} \in W$. Similarly, the right-tail of the distribution of T is obtained as

$$\Pr(T \geq t) = \sum_{T(\tilde{w}) \geq t} h(\tilde{w}) \qquad \textbf{Equation 6.6}$$

The probability distribution of T and its tail areas are obtained in Exact Tests by fast numerical algorithms. In large samples, you can obtain an asymptotic approximation for Equation 6.6. Different approximations apply to the different tests in this chapter and are discussed in the section dealing with the specific tests.

P Value Calculations

The p value is the probability, under H_0, of obtaining a value of the test statistic at least as extreme as the one actually observed. This probability is computed as the tail area of the null distribution of the test statistic. The choice of tail area, left-tail, right-tail, or two-tails, depends on whether you are interested in a one- or two-sided p value, and also on the type of alternative hypothesis you want to detect. The three statistical tests discussed in this chapter are all different in this respect. For the Mann-Whitney test, both one- and two-sided p values are defined, and they are computed as left, right, or two-tailed probabilities, depending on the alternative hypothesis. For the Kolmogorov-Smirnov test, the p values are computed from the right tail as two-sided p values, depending on how the test statistic is defined. Finally, for the Wald-Wolfowitz runs test, only two-sided p values exist, and they are always computed from the left tail of the null distribution of the test statistic. Because of these complexities, it is more useful to define the p value for each test when the specific test is discussed.

Mann-Whitney Test

The Mann-Whitney test is one of the most popular nonparametric two-sample tests. Indeed, the original paper by Frank Wilcoxon (1945), in which this test was first presented, is one of the most widely referenced statistical papers of all time. For a detailed discussion of this test, see Lehmann (1975). It is assumed that sample 1 consists of n_1 observations drawn from the distribution F_1 and that sample 2 consists of n_2 observations drawn for the distribution F_2. The null hypothesis is given by Equation 6.2. The Wilcoxon test is especially suited to detecting departures from the null hypothesis, in which F_2 is shifted relative to F_1 according to the alternative hypothesis

$$H_1: F_2(v) = F_1(v - \theta) \qquad \textbf{Equation 6.7}$$

The shift parameter θ is unknown. If it can be specified a priori that θ must be either positive or negative, the test is said to be one-sided, and a one-sided p value can be used to decide whether to reject H_0. On the other hand, when it is not possible to specify a priori what the sign of θ ought to be, the test is said to be two-sided. In that case, the two-sided p value is used to decide if H_0 can be rejected.

Before specifying how the one- and two-sided p values are computed, the test statistic $T(\tilde{w}) \equiv T$ must be defined. The first step is to replace the raw data, u, by corresponding scores, w. For the Mann-Whitney test, the score, w_{ij}, replacing the original observation, $u_{i,j}$, is simply the rank of that $u_{i,j}$ in the pooled sample of $N = n_1 + n_2$ observations. If there are no ties among the $u_{i,j}$'s, the N ranks thus substituted into the one-way layout will simply be some permutation of the first N integers. If there are ties in the data, however, use mid-ranks instead of ranks.

In order to define the mid-ranks formally, let $a_{[1]} \le a_{[2]} \le \ldots \le a_{[N]}$ denote the pooled sample of all of the N observations in u, represented as a single row of data sorted in ascending order. To allow for the possibility of ties, let there be g distinct observations among the sorted $a_{[i]}$'s with e_1 distinct observations being equal to the smallest value, e_2 distinct observations being equal to the second smallest value, e_3 distinct observations being equal to the third smallest value, and so forth, until finally e_g distinct observations are equal to the largest value. It is now possible to define the mid-ranks precisely. For $l = 1, 2, \ldots g$, the distinct mid-rank assumed by all the e_l observations tied in the lth smallest position is $w_l^* = e_1 + e_2 + \ldots + e_{l-1} + (e_l + 1)/2$.

Finally, you can determine the $a_{[i]}$, and hence the corresponding u_{ij}, with which each w_l^* is associated. You can then substitute the appropriate w_l^* in place of the u_{ij} in the one-way layout u. In this manner you replace u, the original one-way layout of raw data, with w, the corresponding one-way layout of mid-ranks, whose individual elements, w_{ij}, are the appropriate members of the set of the g distinct mid-ranks $(w_1^*, w_2, \ldots w_g)$. The set W of all possible permutations w is defined by Equation 6.3.

The Wilcoxon rank-sum test statistic for the first column (or sample), $T(\tilde{w}) \equiv T$, is defined as the sum of mid-ranks of the first column (or sample) in the two-way layout, $\tilde{w}$. That is, for any $\tilde{w} \in W$,

$$T = \sum_{i=1}^{n_1} \tilde{w}_{ij} \qquad \textbf{Equation 6.8}$$

Its mean is

$$E(T) = n_1(n_1 + n_2 + 1)/2 \qquad \textbf{Equation 6.9}$$

its variance is

$$\text{var}(T) = \frac{n_1 n_2}{12}\left[n_1 + n_2 + 1 - \frac{\Sigma_{l=1}^{g} e_l(e_l^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right] \qquad \textbf{Equation 6.10}$$

and its observed value is

$$t = \sum_{i=1}^{n_1} w_{ij} \qquad \textbf{Equation 6.11}$$

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:

$$u = t - n_1(n_1 + 1)/2 \qquad \textbf{Equation 6.12}$$

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 $\theta > 0$, while small values of *T* indicate a departure from the null hypothesis in the opposite direction, $\theta < 0$. 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

$$p_1 = \min\{\Pr(T \geq t), \Pr(T \leq t)\} \qquad \textbf{Equation 6.13}$$

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

$$p_2 = \Pr(|T - E(T)| \geq |t - E(T)|) \qquad \textbf{Equation 6.14}$$

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

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

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

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

Next, let

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

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

$$CI = \hat{p}_2 \pm 2.576\hat{\sigma}/\sqrt{M} \qquad \textbf{Equation 6.18}$$

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 $\hat{\sigma}$ 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 $\alpha\%$ confidence interval for the exact p value is

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

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

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

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

$$\tilde{p}_1 = \min\{\Phi((t - E(T))/\sigma_T), 1 - \Phi((t - E(T))/\sigma_T)\} \qquad \textbf{Equation 6.21}$$

and the asymptotic two-sided p value is defined as

$$\tilde{p}_2 = 2\tilde{p}_1 \qquad \textbf{Equation 6.22}$$

where $\Phi(z)$ is the tail area to the left of z from a standard normal distribution, and σ_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
10	90	Control
11	108	Control
12	94	Control
13	78	Control
14	105	Control
15	88	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 the control group at the 5% level of significance.

Although it is not necessary for this small data set, you can compute the Monte Carlo estimate of the exact p value. The results of the Monte Carlo analysis, based on 10,000 random permutations of the original one-way layout, are displayed in Figure 6.3.

Figure 6.3 Monte Carlo results for diastolic blood pressure data

Test Statistics[1]

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)	Exact Sig. [2*(1-tailed Sig.)]	Monte Carlo Sig. (2-tailed)			Monte Carlo Sig. (1-tailed)		
						Sig.	99% Confidence Interval		Sig.	99% Confidence Interval	
							Lower Bound	Upper Bound		Lower Bound	Upper Bound
Diastolic Blood Pressure	9.000	75.000	-1.720	.085	.104[2]	.102[3]	.094	.110	.056[3]	.050	.062

[1.] Grouping Variable: Treatment Group
[2.] Not corrected for ties.
[3.] Based on 10000 sampled tables with starting seed 2000000.

Observe that the Monte Carlo estimate, 0.056, agrees very closely with the exact p value of 0.054. Now observe that with 10,000 Monte Carlo samples, the exact p value is contained within the limits (0.050, 0.062) with 99% confidence. Since the threshold p value, 0.05, falls on the boundary of this interval, it appears that 10,000 Monte Carlo samples are insufficient to conclude that the observed result is not statistically significant. Accordingly, to confirm the exact results, you can next perform a Monte Carlo analysis with 30,000 permutations of the original one-way layout. The results are shown in Figure 6.4. This time, the 99% confidence interval is much tighter and does indeed confirm with 99% confidence that the exact p value exceeds 0.05.

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

Test Statistics[1]

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)	Exact Sig. [2*(1-tailed Sig.)]	Monte Carlo Sig. (2-tailed)			Monte Carlo Sig. (1-tailed)		
						Sig.	99% Confidence Interval		Sig.	99% Confidence Interval	
							Lower Bound	Upper Bound		Lower Bound	Upper Bound
Diastolic Blood Pressure	9.000	75.000	-1.720	.085	.104[2]	.102[3]	.098	.107	.056[3]	.053	.059

1. Grouping Variable: Treatment Group
2. Not corrected for ties.
3. Based on 3000 sampled tables with starting seed 20000000.

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 F_1 and F_2 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

$$H_2\colon\ F_2(v) \neq F_1(v), \text{ for at least one value of } v \qquad \textbf{Equation 6.23}$$

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 F_1 and the empirical CDF for F_2. "Statistical Methods" on p. 76 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 $w = u$ 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 u_{ij}'s. Now let $(u_{[11]} \le u_{[21]} \le \ldots \le u_{[n_1 1]})$ denote the observations from the first sample sorted in ascending order, and let $(u_{12} \le u_{22} \le \ldots \le u_{n_2 2})$ 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 n_j order statistics in the jth sample, $j = 1, 2$, let there be $g_j \le n_j$ distinct order statistics, with e_{1j} observations all tied for first place, e_{2j} observations all tied for second place, and so on until finally, $e_{g_j j}$ observations are all tied for last place. Obviously, $e_{1j} + e_{2j} + \ldots + e_{g_j j} = n_j$. Let $(u^*_{1j} < u^*_{2j} < \ldots < u^*_{g_j j})$ represent the g_j distinct order statistics of sample $j = 1, 2$. You can now compute the empirical CDF's, F_1 for F_1 and F_2 for F_2, as shown below. For $j = 1, 2$, define

$$\hat{F}_j(u) = \begin{cases} 0 & \text{if } u < u^*_{[1j]} \\ (e_{[1j]} + e_{[2j]} + \ldots + e_{kj})/n_j & \text{if } u_{kj} \le u < u_{k+1,j} \text{ for } k = 1, 2, \ldots, g_j - 1 \\ 1 & \text{if } u \ge u^*_{g_j, j} \end{cases}$$

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

$$Z = T(\sqrt{n_1 n_2/(n_1 + n_2)}) \qquad \textbf{Equation 6.24}$$

where T is defined as

$$T = \max_v \left[\left|\hat{F}_1(v) - \hat{F}_2(v)\right|\right] \qquad \textbf{Equation 6.25}$$

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

$$p_2 = \Pr(T \ge t) \qquad \textbf{Equation 6.26}$$

When the exact p value is too difficult to compute, you can resort to Monte Carlo sampling. The Monte Carlo estimate of p_2 is denoted by $\hat{p}_2$. It is computed as shown below:

1. Generate a new one-way layout of scores by permuting the original layout of raw data, u, 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

$$Z = \begin{cases} 1 & \text{if } T \geq t \\ 0 & \text{otherwise} \end{cases}$$ **Equation 6.27**

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 6.28**

Next, let

$$\hat{\sigma} = \left[\frac{1}{M-1} \sum_{l=1}^{M} (z_1 - \hat{p}_2)^2 \right]^{1/2}$$ **Equation 6.29**

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

$$CI = \hat{p}_2 \pm 2.576 \hat{\sigma} / \sqrt{M}$$ **Equation 6.30**

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 σ 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

$$CI = [0, 1 - (1 - \alpha/100)^{1/M}]$$ **Equation 6.31**

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

$$CI = [(1 - \alpha/100)^{1/M}, 1]$$ **Equation 6.32**

Exact Tests uses default values of M=10000 and α=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.

The asymptotic two-sided p value, $\hat{p}_2$, is based on the following limit theorem:

$$\lim_{n_1, n_2 \to \infty} \Pr(\sqrt{n_1 n_2 / (n_1 + n_2)}\, T \le z) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2} \qquad \textbf{Equation 6.33}$$

Although the right side of Equation 6.33 has an infinite number of terms, in practice you need to compute only the first few terms of the above expression before convergence is achieved.

Example: Effectiveness of Vitamin C

These data are taken from Lehmann (1975). The effectiveness of vitamin C in orange juice and synthetic ascorbic acid was compared in 20 guinea pigs (divided at random into two groups). Figure 6.5 shows the data displayed in the Date Editor. There are two variables in these data—*score* represents the results, in terms of length of odontoblasts (rounded to the nearest integer) after six weeks; *source* indicates the source of the vitamin C, either *orange juice* or *ascorbic acid*.

Figure 6.5 Effectiveness of vitamin C in orange juice and ascorbic acid

	score	source
1	8	Orange Juice
2	8	Orange Juice
3	10	Orange Juice
4	10	Orange Juice
5	10	Orange Juice
6	15	Orange Juice
7	15	Orange Juice
8	16	Orange Juice
9	18	Orange Juice
10	22	Orange Juice
11	4	Ascorbic Acid
12	5	Ascorbic Acid
13	6	Ascorbic Acid
14	6	Ascorbic Acid
15	7	Ascorbic Acid
16	7	Ascorbic Acid
17	10	Ascorbic Acid
18	11	Ascorbic Acid
19	11	Ascorbic Acid
20	12	Ascorbic Acid

The results of the two-sample Kolmogorov-Smirnov test for these data are shown in Figure 6.6.

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

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

against the alternative hypothesis

$$H_2: \; F_1(v) \neq F_2(v) \text{ for at least one } v \qquad \textbf{Equation 6.35}$$

The test is completely general, in the sense that no distributional assumptions need to be made about F_1 and F_2. 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 $N = n_1 + n_2$ observations in ascending order, and position them in a single row represented as $(a_{[1]} \leq a_{[2]} \leq \ldots \leq a_{[N]})$.
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 number of runs. On the other hand, if observations from F_1 tend to be smaller than those from F_2, you expect the sorted list to lead with the sample 1 observations and be followed by the sample 2 observations. In the extreme case, there will be only two runs. Likewise, if the observations from F_2 tend to be smaller than those from F_1, you expect the sorted list to lead with the sample 2 observations and be followed by the sample 1 observations. Again, in the extreme case, there will be only two runs. These considerations imply that the p value for testing H_0 against the omnibus alternative H_1 should be the left tail of the random variable, T, at the observed number of runs, t. That is, the exact p value is given by

$$p_1 = \Pr(T \leq t) \qquad \textbf{Equation 6.36}$$

The distribution of T is obtained by permuting the observed one-way layout in all possible ways and assigning the probability (see Equation 6.4) to each permutation. You can also derive this distribution theoretically using the same reasoning that was used in "Runs Test" on p. 51 in Chapter 4; the Monte Carlo p value, $\tilde{p}_1$, and the asymptotic p value, $\tilde{p}_1$, can be obtained similarly, using the results described in this section.

Example: Discrimination against Female Clerical Workers

The following example uses a subset of data published by Gastwirth (1991). In November, 1983, a female employee of Shelby County Criminal Court filed a charge of discrimination in pay between similarly qualified male and female clerical workers.

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

number of runs after rearranging the ties in all possible ways. This is denoted by $t_{\max}$. Exact Tests produces two p values,

$$p_{1,\min} = \Pr(T \leq t_{\min}) \qquad \textbf{Equation 6.37}$$

and

$$p_{1,\max} = \Pr(T \leq t_{\max}) \qquad \textbf{Equation 6.38}$$

Conservative decisions are usually made with $p_{1,\max}$. For the clerical workers data set, the output of the Wald-Wolfowitz test is shown in Figure 6.8.

Figure 6.8 Wald-Wolfowitz runs test for clerical workers data

Frequencies

			N
Starting Monthly Salary	Gender of Worker	Male	3
		Female	6
		Total	9

Test Statistics[1,2]

		Number of Runs	Z	Exact Sig. (1-tailed)	Point Probability
Starting Monthly Salary	Minimum Possible	2[3]	-2.041	.024	.024
	Maximum Possible	4[3]	-.408	.345	.238

1. Wald-Wolfowitz Test
2. Grouping Variable: Gender of Worker
3. There are 1 inter-group ties involving 2 cases.

When ties are broken in all possible ways, the minimum number of runs is 2, and the maximum is 4. The smallest possible exact p value is thus $p_{1,\min} = 0.024$. The largest possible exact p value is $p_{1,\max} = 0.345$. In the interest of being as conservative as possible, this is clearly the one to report. It implies that you cannot reject the null hypothesis that $F_1 = F_2$.

Median Test

The two-sample version of the median test is identical in every respect to the k-sample version discussed in Chapter 8. Please refer to the discussion of the median test in Chapter 8 and substitute $K = 2$ if there are only two samples.

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 $K \times 1$ 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 matched sets of data, where the matching is based on demographic, social, medical or other factors that are a priori known to influence response and are not, therefore, under investigation.
- There are K distinct judges, all evaluating the same set of N applicants and assigning ordinal scores to them.

Many other possibilities exist for generating K related samples of data. In all of these settings, the objective is to determine if the K populations from which the data arose are the same. Tests of this hypothesis are often referred to as **blocked comparisons** to emphasize that the data consist of N independent blocks with K dependent observations within each block. Exact Tests provides three tests for this problem: Friedman's, Cochran's Q, and Kendall's W, also known as Kendall's coefficient of concordance.

Available Tests

Table 7.1 shows the available tests for related samples, the procedure from which they can be obtained, and a bibliographical reference for each test.

Table 7.1 Available tests

Test	Procedure	Reference
Friedman's test	Nonparametric Tests: Tests for Several Related Samples	Lehmann (1975)
Kendall's *W* test	Nonparametric Tests: Tests for Several Related Samples	Conover (1975)
Cochran's *Q* test	Nonparametric Tests: Tests for Several Related Samples	Lehmann (1975)

When to Use Each Test

Friedman's test. Use this test to compare *K* related samples of data. Each observation consists of a $1 \times K$ vector of correlated values, and there are *N* such observations, thus forming an $N \times K$ two-way layout.

Kendall's W test. This test is completely equivalent to Friedman's test. The only advantage of this test over Friedman's is that Kendall's *W* has an interpretation as the coefficient of concordance, a popular measure of association. (See also Chapter 14).

Cochran's Q test. This test is identical to Friedman's test but is applicable only to the special case where the responses are all binary.

Statistical Methods

The observed data for all of the tests in this chapter are represented in the form of a two-way layout, shown in Table 7.2.

Table 7.2 Two-way layout for K related samples

Block	Treatments			
Id	**1**	**2**	**...**	**K**
1	u_{11}	u_{12}	...	u_{lK}
2	u_{21}	u_{22}	...	u_{2K}
.	.	.	...	.
.	.	.		.
.	.	.		.
N	u_{N1}	u_{N2}	...	u_{NK}

This layout consists of N independent blocks of data with K correlated observations within each block. The data are usually continuous (possibly with ties). However, for the Cochran's Q test, the data are binary. Various test statistics can be defined on this two-way layout. Usually, however, these test statistics are defined on ranked data rather than on the original raw data. Accordingly, first replace the K observations, $(u_{i1}, u_{i2}, \ldots u_{iK})$ in block i with corresponding ranks, $(r_{i1}, r_{i2}, \ldots r_{iK})$. If there were no ties among these u_{ij}s , you would assign the first K integers $(1, 2, \ldots K)$, 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_{23} to the third smallest, and so on. Similarly, the K observations in the second block take on e_2 distinct values, with d_{21} of the observations being equal to the smallest value, d_{22} to the next smallest, d_{23} to the third smallest, and so on. Finally, the K observations in the Nth block take on e_N distinct values, with d_{N1} of the observations being equal to the smallest value, d_{N2} to the next smallest, d_{N3} to the third smallest, and so on. It is now possible to define the mid-ranks precisely. For $i = 1, 2, \ldots N$, the e_i distinct mid-ranks in the ith block, sorted in ascending order, are

$$
\begin{aligned}
r^*_{i1} &= (d_{i1} + 1)/2 \\
r^*_{i2} &= d_{i1} + (d_{i2} + 1)/2 \\
&\cdots \\
r^*_{i,ei} &= d_{1i} + d_{2i} + \ldots + (d_{i,ei} + 1)/2
\end{aligned}
$$

Equation 7.1

You can now replace the original observations, $(u_{i1}, u_{i2}, \ldots u_{iK})$, in the ith block with corresponding mid-ranks, $(r_{i1}, r_{i2} \ldots r_{iK})$, where each r_{ij} is the appropriate selection from the set of distinct mid-ranks $(r^*_{i1} < r^*_{i2} < \ldots < r^*_{i,ei})$. 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**	…	***K***
1	r_{11}	r_{12}	…	r_{1K}
2	r_{21}	r_{22}	…	r_{2K}
.	.	.	…	.
.	.	.		.
.	.	.		.
.				
N	r_{N1}	r_{N2}	…	r_{NK}

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

Table 7.4 Two-way layout with two blocks of raw data

Block	Treatments				
ID	1	2	3	4	5
1	1.3	1.1	1.1	1.6	1.1
2	1.9	1.7	1.9	1.9	1.7

For the first block, $e_1 = 3$, with $d_{11} = 3$, $d_{12} = 1$, $d_{13} = 1$. Using Equation 7.1, you can obtain mid-ranks $r^*_{11} = 2$, $r^*_{12} = 4$, and $r^*_{13} = 5$. For the second block, $e_2 = 2$, with $d_{21} = 2$, $d_{22} = 3$. Thus, you obtain mid-ranks $r^*_{21} = 1.5$ and $r^*_{21} = 4$. You can now use these mid-ranks to replace the original u_{ij} values with corresponding r_{ij} values. The modified two-way layout, in which raw data have been replaced by mid-ranks, is displayed as Table 7.5.

Table 7.5 Sample two-way layout with raw data replaced by mid-ranks

Block	Treatments				
ID	1	2	3	4	5
1	4	2	2	5	2
2	4	1.5	4	4	1.5

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

$$r_j = \sum_{i=1}^{N} r_{ij} \qquad \textbf{Equation 7.2}$$

the average rank-sum for treatment j as

$$r_{.j} = \frac{r_j}{N} \qquad \textbf{Equation 7.3}$$

and the average rank-sum across all treatments as

$$r_{..} = \frac{\sum_{j=1}^{K} r_{.j}}{K} = \frac{K+1}{2} \qquad \textbf{Equation 7.4}$$

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

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

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

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

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.

2. Compute the value of the test statistic T for the new two-way layout. Define the random variable

$$Z = \begin{cases} 1 \text{ if } T \geq t \\ 0 \text{ otherwise} \end{cases} \qquad \textbf{Equation 7.8}$$

3. Repeat steps 1 and 2 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

$$\hat{p}_2 = \frac{\sum_{l=1}^{M} z_l}{M} \qquad \textbf{Equation 7.9}$$

Next, let

$$\hat{\sigma} = \left[\frac{1}{M-1} \sum_{l=1}^{M} (z_l - \hat{p}_2)^2 \right]^{1/2} \qquad \textbf{Equation 7.10}$$

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

$$CI = \hat{p}_2 \pm 2.576\hat{\sigma}/\sqrt{M} \qquad \textbf{Equation 7.11}$$

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 σ 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

$$CI = [0, 1 - (1 - \alpha/100)^{1/M}] \qquad \textbf{Equation 7.12}$$

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

$$CI = [(1 - \alpha/100)^{1/M}, 1] \qquad \textbf{Equation 7.13}$$

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.

The asymptotic p value is obtained by noting that the large-sample distribution of T is chi-square with $K-1$ degrees of freedom. Thus, the asymptotic two-sided p value is

$$\tilde{p}_2 = \chi^2_{K-1} \geq t \qquad \textbf{Equation 7.14}$$

One-sided p values are inappropriate for the tests in this chapter, since they all assume that there is no a priori natural ordering of the K treatments under the alternative hypothesis. Thus, large observed values of T are indicative of a departure from H_0 but not of the direction of the departure.

Friedman's Test

The methods discussed in this and succeeding sections of this chapter apply to both the randomization and population models for generating the data. If you assume that the assignment of the treatments to the K subjects within each block is random (the randomized block design), you need make no further assumptions concerning any particular population model for generating the u_{ij}'s. This is the approach taken by Lehmann (1975). However, sometimes it is useful to specify a population model, since it allows you to define the null and alternative hypotheses precisely. Accordingly, following Hollander and Wolfe (1973), you can take the model generating the original two-way layout (see Table 7.2) to be

$$U_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \qquad \textbf{Equation 7.15}$$

for $i = 1, 2, \ldots N$, and $j = 1, 2, \ldots K$, where μ is the overall mean, β_i is the block effect, τ_j is the treatment effect, and the ε_{ij}'s are identically distributed unobservable error terms from an unknown distribution, with a mean of 0. All of these parameters are unknown, but for identifiability you can assume that

$$\sum_{i=1}^{N} \beta_i = \sum_{j=1}^{K} \tau_j = 0$$

Note that U_{ij} is a random variable, whereas u_{ij} is the specific value assumed by it in the data set under consideration. The null hypothesis that there is no treatment effect may be formally stated as

$$H_0: \tau_1 = \tau_2 = \ldots = \tau_K \qquad \textbf{Equation 7.16}$$

Friedman's test has good power against the alternative hypothesis

$$H_1: \ \tau_{j_1} \neq \tau_{j_2} \text{ for at least one} (j_1, j_2) \text{ pair} \qquad \textbf{Equation 7.17}$$

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.

$$T_F = \frac{12\sum_{j=1}^{K}(r_j - Nr_{..})^2}{NK(K+1) - (K-1)^{-1}\sum_{i=1}^{N}\left[\sum_{j=1}^{e_i} d_{ij}^3 - K\right]} \qquad \textbf{Equation 7.18}$$

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 *agitate* give the subjects's skin measurements (adjusted for initial level) in millivolts for each of the emotions studied.

Figure 7.1 Effect of hypnosis on skin potential

subject	fear	happy	depress	calmness	agitate
1	23	58	11	24	34
2	23	52	10	20	40
3	23	54	22	21	22

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.

Figure 7.3 Monte Carlo 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	Chi-Square	df	Asymp. Sig.	Monte Carlo Sig.		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
3	9.153	4	.057	.027	.023	.032

1. Friedman Test

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,* or coefficient of concordance, was actually developed as a measure of association, with the N blocks representing N independent judges, each one assigning ranks to the same set of K applicants (Kendall and Babington-Smith, 1939). Kendall's W measures the extent to which the N judges agree on their rankings of the K applicants.

Kendall's W bears a close relationship to Friedman's test; Kendall's W is in fact a scaled version of Friedman's test statistic:

$$W = \frac{T_F}{N(K-1)} \qquad \textbf{Equation 7.19}$$

The exact permutation distribution of W is identical to that of T_F, and tests based on either W or T_F produce identical p values. The scaling ensures that $W = 1$ if there is perfect agreement among the N judges in terms of how they rank the K applicants. On the other hand, if there is perfect disagreement among the N judges, $W = 0$. The fact that the judges don't agree implies that they don't rank the K applicants in the same order. So each applicant will fare well at the hands of some judges and poorly at the hands of others. Under perfect disagreement, each applicant will fare the same overall and will thereby produce an identical value for $R_{\cdot j}$. This common value of $R_{\cdot j}$ will be $R_{\cdot\cdot}$, and as a consequence, $W = 0$.

Example: Attendance at an Annual Meeting

This example is taken from Siegel and Castellan (1988). The Society for Cross-Cultural Research (SCCR), decided to conduct a survey of its membership on factors influencing attendance at its annual meeting. A sample of the membership was asked to rank eight factors that might influence attendance. The factors, or variables, were *airfare*, *climate*, *season*, *people*, *program*, *publicity*, *present*, and *interest*. Figure 7.4 displays the data in the Data Editor and shows how three members (raters 4, 21, and 11) ranked the eight variables.

Figure 7.4 Rating of factors affecting decision to attend meeting

id	airfare	climate	season	people	program	publicty	present	interest
4	5	6	7	1	2	4	3	8
21	1	7	6	2	3	5	4	8
11	4	5	1	3	2	7	6	8

To test the null hypothesis that Kendall's coefficient of concordance is 0, out of the eight possible ranks, each rater (judge) assigns a random rank to each factor (applicant). The results are shown in Figure 7.5.

Figure 7.5 Results of Kendall's W for data on factors affecting decision to attend meeting

Ranks

	Mean Rank
AIRFARE	3.33
CLIMATE	6.00
SEASON	4.67
PEOPLE	2.00
PROGRAM	2.33
PUBLICTY	5.33
PRESENT	4.33
INTEREST	8.00

Test Statistics[1]

N	Kendall's W[1]	Chi-Square	df	Asymp. Sig.	Monte Carlo Sig.		
					Sig.	99% Confidence Interval	
						Lower Bound	Upper Bound
3	.656	13.778	7	.055	.022[2]	.018	.026

1. Kendall's Coefficient of Concordance
2. Based on 10000 sampled tables with starting seed 2000000.

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 on 10,000 random permutations of the original two-way layout of mid-ranks. The Monte Carlo estimate is 0.022, less than half of the asymptotic p value, and is strongly suggestive that the coefficient of concordance is not 0. The 99% confidence interval for the exact p value is (0.022, 0.026). It confirms that you can reject the null hypothesis that there is no association at the 5% significance level, since you are 99% assured that the exact p value is no larger than 0.026.

Equation 7.19 implies that Friedman's test and Kendall's W test will yield identical p values. This can be verified by running Friedman's test on the data shown in Figure 7.4. Figure 7.6 shows the asymptotic and Monte Carlo p values for Friedman's test and demonstrates that they are the same as those obtained with Kendall's W test. The Monte Carlo equivalence was achieved by using the same starting seed and the same number

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.

Figure 7.6 Friedman's test results for data on factors affecting decision to attend meeting

Test Statistics[1]

N	Chi-Square	df	Asymp. Sig.	Monte Carlo Sig.		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
3	13.778	7	.055	.022	.018	.026

1. Friedman Test

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 $N = 2$ judges, each ranking K applicants. Could this measure be extended if N exceeded 2? One approach might be to form $N!/(2!(N-2)!)$ distinct pairs of judges. Then each pair would yield a value for Spearman's rank-order correlation coefficient. Let $\text{ave}(R_S)$ denote the average of all these Spearman correlation coefficients. If there are no ties in the data you can show (Conover, 1980) that

$$\text{ave}(R_S) = \frac{NW - 1}{N - 1} \qquad \textbf{Equation 7.20}$$

Thus, the average Spearman rank-order correlation coefficient is linearly related to Kendall's coefficient of concordance, and you have a natural way of extending the concept correlation from a measure of association between two judges to one between several judges.

This can be illustrated with the data in Figure 7.4. As already observed, Kendall's W for these data is 0.656. Using the procedure discussed in "Spearman's Rank-Order Correlation Coefficient" on p. 174 in Chapter 14, you can compute Spearman's correlation coefficient for all possible pairs of raters. The Spearman correlation coefficient between rater 4 and rater 21 is 0.7381. Between rater 4 and rater 11, it is 0.2857. Finally, between rater 21 and rater 11, it is 0.4286. Therefore, the average of the three Spearman correlation coefficients is $(0.7381 + 0.2857 + 0.4286)/3 = 0.4841$. Substituting $N = 3$ and $W = 0.6561$ into Equation 7.20, you also get 0.4841.

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

$$\log \frac{\pi_{ij}}{1-\pi_{ii}} = \mu + \beta_i + \tau_j \qquad \textbf{Equation 7.21}$$

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

$$\sum_{i=1}^{N} \beta_i = \sum_{j=1}^{K} \tau_j = 0$$

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

$$H_0: \ (\tau_1 = \tau_2 = \ldots = \tau_K) \qquad \textbf{Equation 7.22}$$

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

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

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

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

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 $(B_1 + B_2 + \ldots + B_K)/K$. 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 $(u_{ij} = 0)$. Figure 7.7 shows the data displayed in the Data Editor.

Figure 7.7 Crossover clinical trial of analgesic efficacy

id	placebo	aspirin	drug
1	Failure	Success	Success
2	Failure	Success	Success
3	Success	Failure	Success
4	Failure	Failure	Failure
5	Failure	Failure	Success
6	Failure	Success	Success
7	Success	Failure	Success
8	Failure	Failure	Success
9	Failure	Failure	Failure
10	Failure	Failure	Success
11	Failure	Success	Failure
12	Failure	Failure	Success

The Cochran's Q test can be used to determine if the response rates for the three treatments differ. The results are displayed in Figure 7.8.

Figure 7.8 Cochran's Q results for study of analgesic efficacy

Frequencies

	Value	
	0	1
Placebo	10	2
Aspirin	8	4
New Drug	3	9

Test Statistics[1]

N	Cochran's Q	df	Asymp. Sig.	Exact Sig.	Point Probability
12	7.800[1]	2	.020	.026	.019

1. 0 is treated as a success.

The exact p value is 0.026 and indicates that the three treatments are indeed significantly different at the 5% level. The asymptotic p value, 0.020, confirms this result. In this data set, there was very little difference between the exact and the asymptotic inference. However, the data set is fairly small, and a slightly different data configuration could have resulted in an important difference between the exact and asymptotic p values. To illustrate this point, ignore the data provided by the 12th subject. Running Cochran's Q test once more, this time on only the first 11 subjects, yields the results shown in Figure 7.9.

Figure 7.9 Cochran's Q results for reduced analgesic efficacy data

Frequencies

	Value	
	0	1
Placebo	9	2
Aspirin	7	4
New Drug	3	8

Test Statistics[1]

N	Cochran's Q	df	Asymp. Sig.	Exact Sig.	Point Probability
11	6.222[1]	2	.045	.059	.024

1. 0 is treated as a success.

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

This chapter deals with tests based on K independent samples of data drawn from K distinct populations. The objective is to test the null hypothesis that the K populations all have the same response distributions against the alternative that the response distributions are different. The data could also arise from randomized clinical trials in which each subject is assigned, according to a prespecified randomization rule, to one of K treatments. Here it is not necessary to make any assumptions about the underlying populations from which these subjects were drawn, and the goal is simply to test that the K treatments are the same in terms of the responses they produce. Lehmann (1975) has demonstrated clearly that the same statistical methods are applicable whether the data arose from a population model or a randomization model. Thus, no distinction will be made between the two ways of gathering the data.

This chapter generalizes the tests for two independent samples, discussed in Chapter 6, to tests for K independent samples. There are two important distinctions between the structure of the data in this chapter and in Chapter 7 (the chapter on K related samples). In this chapter, the data are independent both within a sample and across samples; in Chapter 7, the data are correlated across the K samples. Also, in this chapter, the sample sizes can differ across the K samples, with n_j being the size of the jth sample; in Chapter 7, the sample size, N, is required to be the same for each of the K samples.

Available Tests

Table 8.1 shows the available tests for several independent samples, the procedure from which they can be obtained, and a bibliographical reference for each test.

Table 8.1 Available tests

Tests	Commands	References
Median test	Nonparametric Tests: Tests for Several Independent Samples	Gibbons (1985
Kruskal-Wallis Test	Nonparametric Tests: Tests for Several Independent Samples	Siegel & Castellan (1988)
Jonckheere-Terpstra Test	Nonparametric Tests: Tests for Several Independent Samples	Hollander & Wolfe (1973)

The Kruskal-Wallis and the Jonckheere-Terpstra tests are also discussed in the chapters on crosstabulated data. The Kruskal-Wallis test also appears in Chapter 11, which discusses singly-ordered $r \times c$ contingency tables. The Jonckheere-Terpstra test also appears in Chapter 12, which deals with doubly-ordered $r \times c$ contingency tables. These tests are applicable both to data arising from nonparametric continuous univariate-response models (discussed in this chapter) and to data arising from categorical-response models such as the multinomial, Poisson, or hypergeometric models (discussed in later chapters). The tests in the two settings are completely equivalent, although the formulas for the test statistics might differ slightly to reflect the different mathematical models giving rise to the data.

When to Use Each Test

The tests discussed in this chapter are of two broad types: those appropriate for use against unordered alternatives and those for use against ordered alternatives. Following a discussion of these two types of tests, each individual test will be presented, along with the null and alternative hypotheses.

Tests Against Unordered Alternatives

Use the median test or the Kruskal-Wallis test if the alternatives to the null hypothesis of equality of the K populations are unordered. The term **unordered alternatives** means that there can be no a priori ordering of the K populations from which the samples were drawn, under the alternative hypothesis. As an example, the K populations might represent K distinct cities in the United States. Independent samples of individuals are taken from each city and some measurable characteristic, say annual income, is selected as the response. There is no a priori reason why the cities should be arranged in increasing order of the income distributions of their residents, under the alternative hypothesis. All you can reasonably say is that the income distributions are unequal.

For tests against unordered alternatives, the only conclusion you can draw when the null hypothesis is rejected is that the K populations do not all have the same probability distribution. Therefore, a one-sided p value cannot be defined for testing a specific

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 more of the available information by considering the relative magnitude of each observation when compared with every other observation. On the other hand, it is the most general of the available tests, making no assumptions about the alternative hypothesis.

Kruskal-Wallis test. This is one of the most popular nonparametric tests for comparing *K* independent samples. It is the nonparametric analog of one-way ANOVA. In *p* value calculations, mid-ranks are substituted for the raw data and exact permutational distributions are substituted for *F* distributions derived from normality assumptions. It has good power against **location-shift alternatives**, where the distributions from which the samples were drawn have the same general shape but their means are shifted with respect to each other. It is about 98% as efficient as one-way ANOVA for comparing *K* samples when the underlying populations are normal and have a common variance.

Tests Against Ordered Alternatives

Use the Jonckheere-Terpstra test if the alternatives to the null hypothesis of equality of the *K* populations are ordered. The term **ordered alternatives** means that there is a natural a priori ordering of the *K* populations from which the samples were drawn, under the alternative hypothesis. For example, the *K* populations might represent *K* progressively increasing doses of some drug. Here the null hypothesis is that the different dose levels all produce the same response distributions; the alternative hypothesis is that there is a dose-response relationship in which increases in drug dose lead to increases in the magnitude of the response. In this setting, there is indeed an a priori natural ordering of the *K* populations in terms of increased dose levels of the drug. One of the implications of natural ordering under the alternative hypothesis is that the ordering could be either ascending or descending. For the dose-response example, you could define a one-sided *p* value for testing the null hypothesis against the alternative that an increase in drug dose increases the probability of response. But you could also define a one-sided *p* value against the alternative that it leads to a decrease in the probability of response. A two-sided *p* value could be defined to test the null hypothesis against either alternative. Thus, for tests against ordered alternatives, both one- and two-sided *p* values are relevant.

Statistical Methods

The data for all the tests in this chapter consist of K independent samples each of size n_j, $j = 1,2,\ldots K$, where $n_1 + n_2 + \ldots n_K = N$. These N observations can be represented in the form of the one-way layout shown in Table 8.2.

Table 8.2 One-way layout for K independent samples

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

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

$$F_j(v) = \Pr(V \le v | j),\ j = 1,2,\ldots K \qquad \textbf{Equation 8.1}$$

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

$$H_O:\ F_1 = F_2 = \ldots = F_K \qquad \textbf{Equation 8.2}$$

In order to test H_0 by nonparametric methods, it is necessary to replace the original observations in the above 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, as you will see in the individual sections on each test. Let w_{ij} be the score corresponding to u_{ij}. Then the one-way layout, with the original data replaced by scores, is shown in Table 8.3.

Table 8.3 One-way layout with scores replacing original data

Samples			
1	**2**	…	***K***
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

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

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		

As discussed in "Kruskal-Wallis Test" on p. 127, to run the Kruskal-Wallis test on these data, you must replace them with their ranks. The one-way layout of observed scores, with the original data replaced by their ranks, is shown in Table 8.5.

Table 8.5 One-way layout with ranks replacing original data

Samples		
1	**2**	**3**
3.5	6	9
5	1	10
7	3.5	11
8		
2		

This one-way layout of ranks is denoted by *w*. It is the one actually observed. Notice that two observation 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, use the mid-rank, $(3 + 4)/2 = 3.5$, as the rank for each of them in *w*. The symbol *W* represents the set of all possible one-way layouts in which entries are the 11 numbers in *w*, with 5 numbers in column 1, 3 numbers in column 2, and 3 numbers in column 3. Thus, *w* is one member of *W*. (It is the one actually observed.) Another member is w', where w' is a different permutation of the numbers in *w,* as shown in Table 8.6.

Table 8.6 Permutation of the observed one-way layout of scores

Sample		
1	**2**	**3**
6	5	9
1	8	10
3.5	7	11
3.5		
2		

All of the test statistics in this chapter are univariate functions of $\tilde{w} \in W$. Let the test statistic be denoted by $T(\tilde{w}) \equiv T$, and its observed value be denoted by $t(w) \equiv t$. The functional form of $T(\tilde{w})$ will be defined separately for each test in subsequent sections of this chapter. Following is a discussion of the null distribution of *T*—how it can be derived in general, and how it is used for *p* value computations.

Distribution of T

In order to test the null hypothesis, H_0, you need to derive the distribution of T under the assumption that H_O is true. This distribution is obtained by the following permutational argument:

If H_O is true, every member $\tilde{w} \in W$ 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 N subjects to K treatments in accordance with a predetermined randomization rule. Therefore, no distinction will be made between these two ways of gathering the data.

It follows from the above permutational argument that the exact probability of observing any $\tilde{w} \in W$ is

$$h(\tilde{w}) = \frac{\Pi^{K}_{j=1} n_j!}{N!} \qquad \textbf{Equation 8.4}$$

which does not depend on the specific way in which the original one-way layout, w, was permuted. Then

$$\Pr(T = t) = \sum_{T(\tilde{w}) = t} h(\tilde{w}) \qquad \textbf{Equation 8.5}$$

the sum being taken over all $\tilde{w} \in W$. Similarly, the right tail of the distribution of T is obtained as

$$\Pr(T \geq t) = \sum_{T(\tilde{w}) \geq t} h(\tilde{w}) \qquad \textbf{Equation 8.6}$$

The probability distribution of T and its tail areas are obtained in Exact Tests by numerical algorithms. In large samples, you can obtain an asymptotic approximation for Equation 8.6. Different approximations apply to the various tests described in this chapter and are discussed in the sections specific to each test.

P Value Calculations

The p value is the probability, under H_0, of obtaining a value of the test statistic at least as extreme as the one actually observed. The exact, Monte Carlo, and asymptotic p values can be computed for tests on K independent samples as follows.

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

$$p_2 = \Pr(T \geq t) = \sum_{T \geq t} h(\tilde{w}) \qquad \textbf{Equation 8.7}$$

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 H_0 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

$$p_1 = \min\{\Pr(T \geq t), \Pr(T \leq t)\} \qquad \textbf{Equation 8.8}$$

and the exact two-sided p value is defined as

$$p_2 = \Pr(|T - E(T)| \leq |t - E(T)|) \qquad \textbf{Equation 8.9}$$

where $E(T)$ 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 sampling. Below, Monte Carlo sampling is used to estimate the exact p value given by Equation 8.7. The same procedure can be readily adapted to Equation 8.8 and Equation 8.9.

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

$$Z = \begin{cases} 1 \text{ if } T \geq t \\ 0 \text{ otherwise} \end{cases} \qquad \textbf{Equation 8.10}$$

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

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

Next, let

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

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

$$CI = \hat{p}_2 \pm 2.576\hat{\sigma}/\sqrt{M} \qquad \textbf{Equation 8.13}$$

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 $\hat{\sigma}$ is based on inverting an exact binomial hypothesis test when an extreme outcome is encountered. It can be shown that if $\hat{p}_2 = 0$, an $\alpha\%$ confidence interval for the exact p value is

$$CI = [0, 1 - (1 - \alpha/100)^{1/M}] \qquad \textbf{Equation 8.14}$$

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

$$CI = [(1 - \alpha/100)^{1/M}, 1] \qquad \textbf{Equation 8.15}$$

Exact Tests uses default values of $M = 10000$ and $\alpha = 99\%$. 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 $K-1$ degrees of freedom. The asymptotic p value is thus

$$\hat{p}_2 = \Pr(x^2_{K-1} \geq t) \qquad \textbf{Equation 8.16}$$

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

$$\tilde{p}_1 = \min\{\Phi(t - E(T)/\sigma_T, 1 - \Phi(t - E(T)/\sigma_T\} \qquad \textbf{Equation 8.17}$$

and the asymptotic two-sided p value is defined as

$$\tilde{p}_2 = 2\tilde{p}_1 \qquad \textbf{Equation 8.18}$$

where $\Phi(z)$ is the tail area to the left of z from a standard normal distribution, and σ_T is the standard deviation of T. Explicit expressions for $E(T)$ and σ_T are provided in “Jonckheere-Terpstra Test” on p. 131.

Median Test

The median test is a nonparametric procedure for testing the null hypothesis H_0, given by Equation 8.2, against the general alternative

$$H_1\text{: There exists at least one } (j_1, j_2) \text{ pair such that } F_{j1} \neq F_{j2} \qquad \textbf{Equation 8.19}$$

The median test is an omnibus test designed for a very general alternative hypothesis. It requires no assumptions about the K distributions, $F_j, j = 1,2,\ldots K$, being tested. However if you have additional information about these distributions—for example, if you believe that they have the same shape but differ from one another by shift parameters under the alternative hypothesis—there are more powerful tests available.

To define the test statistic for the median test, the first step is to transform the original one-way layout of data, as shown in Table 8.2, into a one-way layout of scores, as shown in Table 8.3. To compute these scores, first obtain the grand median, δ, for the pooled sample of size N. The median is calculated in the following way. Let $\alpha_{[1]} \leq \alpha_{[2]} \ldots \leq \alpha_{[N]}$ be the pooled sample of u_{ij} values, sorted in ascending order. Then

$$\delta = \begin{cases} \alpha_{[(n+1)/2]} & \text{if } N \text{ is odd} \\ (\alpha_{[n/2]} + \alpha_{[(n+2)/2]})/2 & \text{if } N \text{ is even} \end{cases} \qquad \textbf{Equation 8.20}$$

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

$$\delta = \begin{cases} 1 \text{ if } u_{ij} \leq \delta \\ 0 \text{ if } u_{ij} > \delta \end{cases} \qquad \textbf{Equation 8.21}$$

Define

$$w_j = \sum_{i=1}^{n_j} w_{ij} \qquad \textbf{Equation 8.22}$$

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

$$m = \sum_{j=1}^{K} w_j \qquad \textbf{Equation 8.23}$$

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 $2 \times K$ 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**	
≤ Median	w_1	w_2	…	w_K	m
> Median	$n_1 - w_1$	$n_2 - w_2$	…	$n_K - w_K$	$N - m$
Column Total	n_1	n_2	…	n_K	N

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

$$h(w) = \frac{\prod_{j=1}^{K} \binom{n_j}{w_j}}{\binom{N}{m}} \qquad \textbf{Equation 8.24}$$

For any $\tilde{w} \in W$, the test statistic for the median test is the usual Pearson chi-square statistic

$$T = \sum_{j=1}^{K} \frac{(\tilde{w}_j - n_j m/N)^2}{n_j m/N} + \sum_{j=1}^{K} \frac{(n_j - \tilde{w}_j - n_j(N-m)/N)^2}{n_j(N-m)/N} \qquad \textbf{Equation 8.25}$$

Thus, if t is the value of T actually observed, the exact two–sided p value for the median test is given by

$$p_2 = \sum_{T \le t} h(\tilde{w}) \qquad \textbf{Equation 8.26}$$

the sum being taken over all $\tilde{w} \in W$ for which $T(\tilde{w}) \le t$. An asymptotic approximation to p_2 is obtained by noting that T converges to the chi-square distribution with $K-1$ degrees of freedom. Therefore,

$$p_2 = \Pr(x^2_{K-1} \ge t) \qquad \textbf{Equation 8.27}$$

The Monte Carlo two-sided p value is obtained as described in "P Value Calculations" on p. 119. Alternatively, you can generate a sequence of M $2 \times K$ contingency tables, $\tilde{w}_1, \tilde{w}_2, \ldots \tilde{w}_m$, each with the same margins as Table 8.7, such that table $\tilde{w}_1$ is generated with probability $h(\tilde{w}_1)$, given by Equation 8.24. For each table generated in this way, you can compute the test statistic, t_1, and define a quantity $z_1 = 1$ if $t_1 = t$; 0 otherwise. The Monte Carlo estimate of p_2 is

$$\hat{p}_2 = \sum_{l=1}^{M} z_l / M \qquad \textbf{Equation 8.28}$$

The 99% Monte Carlo confidence interval for the true p value is calculated by Equation 8.13.

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
16	4	1
17	4	6
18	4	7
19	4	7
20	4	7
21	4	8
22	4	8
23	4	10
24	5	7
25	5	10
26	5	11
27	5	12
28	5	13

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.

Figure 8.2 Median test results for hematologic toxicity data

Frequencies

		Drug Regimen				
		Drug 1	Drug 2	Drug 3	Drug 4	Drug 5
Days with WBC < 500	> Median	2	1	2	3	4
	<= Median	2	4	3	6	1

Test Statistics[1]

	N	Median	Chi-Square	df	Asymp. Sig.	Exact Sig.	Point Probability
Days with WBC < 500	28	7.00	4.317[2]	4	.365	.429	.037

1. Grouping Variable: Drug Regimen

2. 9 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.7.

Figure 8.3 Monte Carlo median test results for hematologic toxicity data

Test Statistics[1]

	N	Median	Chi-Square	df	Asymp. Sig.	Monte Carlo Sig.		
						Sig.	99% Confidence Interval	
							Lower Bound	Upper Bound
Days with WBC < 500	28	7.00	4.317[2]	4	.365	.432[3]	.419	.444

1. Grouping Variable: Drug Regimen

2. 9 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.7.

3. Based on 10000 sampled tables with starting seed 2000000.

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 drugs produce different distributions for the WBC. The asymptotic p value, 0.365, supports this conclusion, but in this small data set, it is not a good approximation of the exact p value. On the other hand, the Monte Carlo estimate of the exact p value, 0.432, comes much closer to the exact p value. The 99% Monte Carlo

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

		Drug Regimen				
		Drug 1	Drug 2	Drug 3	Drug 4	Drug 5
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.

Figure 8.5 Pearson's chi-square results for hematologic toxicity data, divided by the median

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)
Pearson Chi-Square	4.317[1]	4	.365	.429
N of Valid Cases	28			

1. 9 cells (90.0%) have expected count less than 5. The minimum expected count is 1.71.

Kruskal-Wallis Test

The Kruskal-Wallis test (Siegel and Castellan, 1988) is a very popular nonparametric test for comparing K independent samples. When $K = 2$, 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

$$U_{ij} = \mu + \tau_j + \varepsilon_{ij} \qquad \textbf{Equation 8.29}$$

for all $i = 1,2,\ldots n_j$ and $j = 1,2,\ldots K$. In this model, μ is the overall mean, τ is the treatment effect, and the ε_{ij}'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

$$\sum_{j=1}^{K} \tau_j = 0 \qquad \textbf{Equation 8.30}$$

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

$$H_0: \ \tau_1 = \tau_2 = \ldots = \tau_K \qquad \textbf{Equation 8.31}$$

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

$$H_2: \ \tau_{j1} \neq \tau_{j2} \text{ for at least one } (j_1, j_2) \text{ pair} \qquad \textbf{Equation 8.32}$$

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 τ_j's are ordered under the alternative hypothesis—a more powerful test would be the Jonckheere-Terpstra test (Hollander and Wolfe, 1973), discussed on p. 131.

To define the Kruskal-Wallis test statistic, the first step is to convert the one-way layout, *u*, of raw data, as shown in Table 8.2, into a corresponding one-way layout of scores, *w*, as shown in Table 8.3. The scores, w_{ij}, for the Kruskal-Wallis test are the ranks of the observations in the pooled sample of size *N*. If there were no ties, the set of w_{ij} values in Table 8.3 would simply be some permutation of the first *N* integers. However, to allow for the possibility that some observations might be tied, you can assign the mid-rank of a set of tied observations to each of them. The easiest way to explain how the mid-ranks are computed is by considering a numerical example. Suppose that $u_{13}, u_{17}, u_{21}, u_{32}$ are all tied at the same numerical value, say 55. Assume that these four observations would occupy positions 15, 16, 17, and 18, if all the *N* observations were pooled and then sorted in ascending order. In this case, you would assign the mid-rank $(15 + 16 + 17 + 18)/2 = 16.5$ to these four tied observations. Thus, $w_{13} = w_{17} = w_{21} = w_{32} = 16.5$.

More generally, let $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n$ denote the pooled sample of all of the *N* observations sorted in ascending order. To allow for the possibility of ties, let there be *g* distinct observations among the sorted α_i's, with e_1 distinct observations being equal to the smallest value, e_2 distinct observations being equal to the second smallest value, e_3 distinct observations being equal to the third smallest value, and so on, until, finally, e_g distinct observations are equal to the largest value. It is now possible to define the mid-ranks precisely. For $l = 1, 2, \ldots g$, the distinct mid-rank assumed by all of the e_1 observations tied in the *l*th smallest position is

$$w_l^* = e_1 + e_2 + \ldots + e_{l-1} + (e_l + 1)/2$$

In this way, the original one-way layout of raw data is converted into a corresponding one-way layout of mid-ranks.

Next, for any treatment j, where $j = 1,2,\ldots,K$, define the rank-sum as

$$w_j = \sum_{i=1}^{n_j} w_{ij} \qquad \textbf{Equation 8.33}$$

The Kruskal-Wallis test statistic, $T(\tilde{w}) \equiv T$, for any $\tilde{w} \in W$, can now be defined as

$$T = \frac{12}{N(N+1)[1-(\lambda/(N^3-N))]} \sum_{j=1}^{K} [\tilde{w}_j - n_j(N+1)/2]^2/n_j \qquad \textbf{Equation 8.34}$$

where λ is a tie correction factor given by

$$\lambda = \sum_{l=1}^{g} (e_l^3 - e_l) \qquad \textbf{Equation 8.35}$$

The Kruskal-Wallis test is also defined in Chapter 11, using the notation developed for analyzing $r \times c$ contingency tables. The two definitions are equivalent. Since the test is applicable to both continuous and categorical data, the test statistic is defined twice, once in the context of a one-way layout and once in the context of a contingency table.

Let t denote the value of T actually observed from the data. The exact, Monte Carlo, and asymptotic p values based on the Kruskal-Wallis statistic can be obtained as discussed in "P Value Calculations" on p. 119. The exact two-sided p value is computed as shown in Equation 8.7. The Monte Carlo two-sided p value is computed as in Equation 8.11, and the asymptotic two-sided p value is computed as shown in Equation 8.16. One-sided p values are not defined for tests against unordered alternatives like the Kruskal-Wallis test.

Example: Hematologic Toxicity Data, Revisited

The Kruskal-Wallis test can be used to reconsider the hematologic toxicity data displayed in Figure 8.1. You can once again compare the five drugs to determine if they have significantly different response distributions. This time, however, the test statistic actually takes advantage of the relative rankings of the different observations instead of simply using the information that an observation is either above or below the pooled

median. Thus, you can expect the Kruskal-Wallis test to be more powerful than the median test. Although it is too difficult to obtain the exact p value for this data set, you can obtain an extremely accurate Monte Carlo estimate of the exact p value based on a Monte Carlo sample of size 10,000. The results are shown in Figure 8.6.

Figure 8.6 Monte Carlo results of Kruskal-Wallis test for hematologic toxicity data

Ranks

			N	Mean Rank
Days with WBC < 500	Drug Regimen	Drug 1	4	11.88
		Drug 2	5	7.50
		Drug 3	5	17.70
		Drug 4	9	13.50
		Drug 5	5	22.20
		Total	28	

Test Statistics[1,2]

				Monte Carlo Sig.		
					99% Confidence Interval	
	Chi-Square	df	Asymp. Sig.	Sig.	Lower Bound	Upper Bound
Days with WBC < 500	9.415	4	.052	.038[3]	.033	.043

1. Kruskal-Wallis Test
2. Grouping Variable: Drug Regimen
3. Based on 1000 sampled tables with starting seed 2000000.

As expected, the greater power of the Kruskal-Wallis test leads to a smaller p value than obtained with the median test. There is, however, a difference between the asymptotic inference and the exact inference computed by the Monte Carlo estimate. The Monte Carlo estimate of the exact p value is 0.038 and shows that the exact p value is guaranteed to lie in the range (0.033, 0.043) with 99% confidence. Thus, the null hypothesis can be rejected at the 5% significance level. The asymptotic inference, in contrast, was unable to estimate the true p value with this degree of accuracy. It generated a p value of 0.052, which is not significant at the 5% level.

Jonckheere-Terpstra Test

The Jonckheere-Terpstra test (Hollander and Wolfe, 1973) is more powerful than the Kruskal-Wallis test for comparing K samples against ordered alternatives. Once again, assume that the one-way layout shown in Table 8.2 was generated by the model Equation 8.29. The null hypothesis of no treatment effect is again given by Equation 8.31. This time, however, suppose that the alternative hypothesis is ordered. Specifically, the one-sided alternative might be of the form

$$H_1: \ \tau_1 \leq \tau_2 \leq \ldots \leq \tau_K \qquad \textbf{Equation 8.36}$$

implying that as you increase the index j, identifying the treatment, the distribution of responses shifts to the right. Or else, the one-sided alternative might be of the form

$$H'_1: \ \tau_1 \geq \tau_2 \geq \ldots \geq \tau_K \qquad \textbf{Equation 8.37}$$

implying that as you increase the index j, identifying the treatment, the distribution shifts to the left. The two-sided alternative would state that either H_1 or H_2 is true, without specifying which.

To define the Jonchkeere-Terpstra statistic, the first step, as usual, is to replace the original observations with scores. Here, however, let the score, w_{ij}, be exactly the same as the actual observation, u_{ij}. Then $w = u$ and W, as defined by Equation 8.3, is the set of all possible permutations of the one-way layout of actually observed raw data. Now, for any $\tilde{w} \in W$, you compute $K(K-1)/2$ Mann-Whitney counts (see, for example, Lehmann, 1976,), $\{\lambda_{ab}\}$, $1 \leq a \leq (K-1)$, $(a+1) \leq b \leq K$ as follows. For any (a,b), λ_{ab} is the count of the number of pairs, $(\tilde{w}_{\alpha a}, \tilde{w}_{\beta b})$, which are such that $\tilde{w}_{\alpha a} < \tilde{w}_{\beta b}$ plus half the number of pairs, which are such that $(\tilde{w}_{\alpha a} = \tilde{w}_{\beta b})$. The Jonckheere-Terpstra test statistic, $T(\tilde{w}) \equiv T$, is defined as follows:

$$T = \sum_{a=1}^{K-1} \sum_{b=a+1}^{K} \lambda_{ab} \qquad \textbf{Equation 8.38}$$

The mean of the Jonckheere-Terpstra statistic is

$$E(T) = \frac{N^2 - \sum_{j=1}^{K} n_j^2}{4} \qquad \textbf{Equation 8.39}$$

The formula for the variance is more complicated. Suppose, as in "Kruskal-Wallis Test" on p. 127, that there are g distinct u_{ij}'s among all N observations pooled together, with e_1 distinct observations being equal to the smallest value, e_2 distinct observations being

equal to the second smallest value, e_3 distinct observations being equal to the third smallest value, and so on, until, finally, e_g distinct observations are equal to the largest value. The variance of the Jonckheere-Terpstra statistic is

$$\sigma_T^2 = \frac{1}{72}\left[N(N-1)(2N+5) - \sum_{j=1}^{K} n_j(n_j-1)(2n_j+5) - \sum_{l=1}^{g} e_l(e_l-1)(2e_l+5)\right]$$
$$+ \frac{1}{36N(N-1)(N-2)}\left[\sum_{j=1}^{K} n_j(n_j-1)(n_j-2)\right] \times \left[\sum_{l=1}^{g} e_l(e_l-1)(e_l-2)\right]$$
$$+ \frac{1}{8N(N-1)}\left[\sum_{j=1}^{K} n_j(n_j-1)\right]\left[\sum_{l-1}^{g} e_l(e_l-1)\right]$$

Now, let $t(w) \equiv t$ 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. 119. 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 $\{|t - E(T)| \geq |t - E(T)|\}$ instead of the region $\{T \geq t\}$. The Monte Carlo one-sided *p* value can be similarly defined. The asymptotic distribution of *T* is normal, with mean of $E(T)$ and variance σ_T^2. The asymptotic one- and two-sided *p* values are obtained by Equation 8.17 and Equation 8.18, respectively.

Example: Space-Shuttle O-Ring Incidents Data

Professor Richard Feynman, in his delightful book *What Do You Care What Other People Think?* (1988), recounted at great length his experiences as a member of the presidential commission formed to determine the cause of the explosion of the space shuttle Challenger in 1986. He suspected that the low temperature at takeoff caused the O-rings to fail. In his book, he has published the data on temperature versus the number of O-ring incidents, for 24 previous space shuttle flights. These data are shown in Figure 8.7. There are two variables in the data—*incident* indicates the number of O-ring incidents, and is either *none*, *one*, *two*, or *three*; *temp* indicates the temperature in Fahrenheit.

Figure 8.7 Space-shuttle O-ring incidents and temperature at launch

	incident	temp
1	0	66
2	0	67
3	0	67
4	0	67
5	0	68
6	0	68
7	0	70
8	0	70
9	0	72
10	0	73
11	0	75
12	0	76
13	0	76
14	0	78
14	0	78
15	0	79
16	0	80
17	0	81
18	1	57
19	1	58
20	1	63
21	1	70
22	1	70
23	2	75
24	3	53

The null hypothesis is that the temperatures in the four samples (0, 1, 2, or 3 O-ring incidents) have come from the same underlying population distribution. The one-sided alternative hypothesis is that populations with a higher number of O-ring incidents have their temperature distributions shifted to the right of populations with a lower number of O-ring incidents. The Jonckheere-Terpstra test is superior to the Kruskal-Wallis test for this data set because the populations have a natural ordering under the alternative hypothesis. The results of the Jonckheere-Terpstra test for these data are shown in Figure 8.8.

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

Jonckheere-Terpstra Test[1]

	Number of Levels in O-Ring Incidents	N	Observed J-T Statistic	Mean J-T Statistic	Std. Deviation of J-T Statistic	Std. J-T Statistic	Asymp. Sig. (2-tailed)	Exact Significance (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Temperature (Fahrenheit)	4	24	29.500	65.000	15.902	-2.232	.026	.024	.012	.001

1. Grouping Variable: O-Ring Incidents

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

$$T^* = \frac{T - E(T)}{\sigma_T} \qquad \textbf{Equation 8.40}$$

whose observed value is

$$t^* = \frac{t - E(T)}{\sigma_T} \qquad \textbf{Equation 8.41}$$

The output shows that $t = 29.5$, $E(T) = 65$, and $\sigma_T = 15.9$. Therefore, $t^* = -2.232$. The exact one-sided p value is

$$p_1 = \min\{\Pr(T^* \geq t^*), \Pr(T^* \leq t^*)\} \qquad \textbf{Equation 8.42}$$

The exact two-sided p value is

$$p_2 = \Pr(|T^*| \geq |t^*|) \qquad \textbf{Equation 8.43}$$

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 T^* 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 significant correlation between temperature and number of O-ring incidents. The sign of the standardized test statistic, $t^* = -2.232$, is negative, thus implying that higher launch temperatures are associated with fewer O-ring incidents. The two-sided p value would be used if you had no a priori reason to believe that the number of O-ring incidents is negatively correlated with takeoff temperature. Here the exact two-sided p value, 0.024, is also statistically significant.

9 Introduction to Tests on R x C Contingency Tables

This chapter discusses hypothesis tests on data that are cross-classified into contingency tables with r rows and c columns. The cross-classification is based on **categorical variables** that may be either **nominal** or **ordered**. Nominal categorical variables take on distinct values that cannot be positioned in any natural order. An example of a nominal variable is *color* (for example, red, green, or blue). In some statistical packages, nominal variables are also referred to as **class variables**, or unordered variables. Ordered categorical variables take on distinct values that can be ordered in a natural way. An example of an ordered categorical variable is *drug dose* (for example, low, medium, or high). Ordered categorical variables can assume numerical values as well (for example, the drug dose might be categorized into 100 mg/m^2, 200 mg/m^2, and 300 mg/m^2). When the number of distinct numerical values assumed by the ordered variable is very large (for example, the weights of individuals in a population), it is more convenient to regard the variable as **continuous** (possibly with ties) rather than categorical. There is considerable overlap between the statistical methods used to analyze continuous data and those used to analyze ordered categorical data. Indeed, many of the same statistical tests are applicable to both situations. However, the probabilistic behavior of an ordered categorical variable is captured by a different mathematical model than that of a continuous variable. For this reason, continuous variables are discussed separately in Part 1.

This chapter summarizes the statistical theory underlying the exact, Monte Carlo, and asymptotic p value computations for all the tests in Chapter 10, Chapter 11, and Chapter 12. Chapter 10 discusses tests for $r \times c$ contingency tables in which the row and column classifications are both nominal. These are referred to as **unordered contingency tables**. Chapter 11 discusses tests for $r \times c$ contingency tables in which the column classifications are based on ordered categorical variables. These are referred to as **singly ordered contingency tables**. Chapter 12 discusses tests for $r \times c$ tables in which both the row and column classifications are based on ordered categorical variables. These are referred to as **doubly ordered contingency tables**.

Table 9.1 shows an observed $r \times c$ contingency table in which x_{ij} is the count of the number of observations falling into row category i and column category j.

Table 9.1 Observed r x c contingency table

Rows	**Col_1**	**Col_2**	…	**Col_c**	**Row_Total**
Row_1	x_{11}	x_{12}	…	x_{1c}	m_1
Row_2	x_{21}	x_{22}	…	x_{2c}	m_2
.	.	.	…	.	.
.	.	.		.	.
.	.	.		.	.
Row_r	x_{r1}	x_{r2}	…	x_{rc}	m_r
Col_Total	n_1	n_2	…	n_c	N

The main objective is to test whether the observed $r \times c$ 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 $r \times c$ 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 value is recommended. Sometimes, however, a data set is too large for the exact p value computations to be feasible. In this case, the Monte Carlo technique, which is easier to compute, is recommended. The Monte Carlo p value is an extremely close approximation to the exact p value and is accompanied by a fairly narrow confidence interval within which the exact p value is guaranteed to lie (at the specified confidence level). Moreover, by increasing the number of Monte Carlo samples, you can make the width of this confidence interval arbitrarily small. Finally, the exact p value is always recommended. For large, well-balanced data sets, the asymptotic p value is not too different from its exact counterpart, but, obviously, you can't know this for the specific data set on hand without also having the exact or Monte Carlo p value available for comparison. In this section, all three p values will be defined. First, you will see how the exact p value is computed. Then, the Monte Carlo and asymptotic p values will be discussed as convenient approximations to the exact p value computation.

To compute the exact p value of the observed $r \times c$ contingency table, it is necessary to:

1. Define a reference set of $r \times c$ tables in which each table has a known probability under the null hypothesis of no row-by-column interaction.
2. Order all the tables in the reference set according to a discrepancy measure (or test statistic) that quantifies the extent to which each table deviates from the null hypothesis.
3. Sum the probabilities of all tables in the reference set that are at least as discrepant as the observed table.

Defining the Reference Set

Throughout this chapter, x will be used to denote the $r \times c$ contingency table actually observed, and y will denote any generic $r \times c$ contingency table belonging to some well-defined reference set of $r \times c$ 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, product multinomial sampling, and Poisson sampling. Under all three schemes, the probability of observing y depends on unknown parameters relating to the individual cells of the $r \times c$ table. The key to exact nonparametric inference is eliminating all nuisance parameters from the distribution of y. This is accomplished by restricting the sample space to the set of all $r \times c$ contingency tables that have the same marginal sums as the observed table x. Specifically, define the reference set:

$$\Gamma = \left\{ \text{y : y is } r \times c; \sum_{j=1}^{c} y_{ij} = m_i; \sum_{i=1}^{r} y_{ij} = n_j \text{ for all } i, j \right\} \qquad \textbf{Equation 9.1}$$

Then, you can show that, under the null hypothesis of no row-by-column interaction, the probability of observing any $y \in \Gamma$ is

$$P(y) = \frac{\Pi_{j=1}^{c} n_j! \Pi_{i=1}^{r} m_i!}{N! \Pi_{j=1}^{c} \Pi_{i=1}^{r} y_{ij}!} \qquad \textbf{Equation 9.2}$$

Equation 9.2, which is free of all unknown parameters, holds for categorical data whether the sampling scheme used to generate y is full multinomial, product multinomial, or Poisson (Agresti, 1990).

The reference set Γ need not be the actual sample space of the data-generating process. In product multinomial sampling, the row sums are fixed by the experimental design, but the column sums can vary from sample to sample. In full multinomial and Poisson sampling, both the row and column sums can vary. Conditioning on row and column sums is simply a convenient way to eliminate nuisance parameters from the expression for $P(y)$, compute exact p values, and thus guarantee that you will be protected from a conditional type 1 error at any desired significance level. Moreover, since the unconditional type 1 error is a weighted sum of conditional type 1 errors, where the weights are the probabilities of the different marginal configuration, the protection from type 1 errors guaranteed by the conditional test carries over to the unconditional setting. The idea of conditional inference to eliminate nuisance parameters was first proposed by Fisher (1925).

Defining the Test Statistic

For statistical inference, each table $y \in \Gamma$ is ordered by a test statistic or discrepancy measure that quantifies the extent to which the table deviates from the null hypothesis of no row-by-column interaction. The test statistic will be denoted by $D(y)$. Large absolute values of D furnish evidence against the null hypothesis, while small absolute values are consistent with it. The functional form of $D(y)$ for each test is given in the chapter specific to each test. Throughout this chapter, the function $D(y)$ will be used to denote a generic test statistic. Specific instances of test statistics will be denoted by their own unique symbols. For example, for the Pearson chi-square test, the generic symbol $D(y)$ is replaced by $CH(y)$, and the test statistic has the functional form of

$$CH(y) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - m_i n_j / N)^2}{m_i n_j / N} \qquad \textbf{Equation 9.3}$$

Exact Two-Sided P Values

The **exact two-sided *p* value** is defined as the sum of null probabilities of all the tables in Γ that are at least as extreme as the observed table x with respect to D. Specifically,

$$p_2 = \sum_{D(y) \geq D(x)} P(y) = \Pr\{D(y) \geq D(x)\} \qquad \textbf{Equation 9.4}$$

For later reference, define the critical region of the reference set:

$$\Gamma^* = \{y \in \Gamma : D(y) \geq D(x)\} \qquad \textbf{Equation 9.5}$$

Computing Equation 9.4 is sometimes rather difficult because the size of the reference set Γ grows exponentially. For example, the reference set of all 5×6 tables with row sums of (7, 7, 12, 4, 4) and column sums of (4, 5, 6, 5, 7, 7) contains 1.6 billion tables. However, the tables in this reference set are all rather sparse and unlikely to yield accurate p values based on large sample theory. Exact Tests uses network algorithms based on the methods of Mehta and Patel (1983, 1986a, 1986b) to enumerate the tables in Γ implicitly and thus quickly identify those in Γ^*. This makes it feasible to compute exact p values for many seemingly intractable data sets such as the one above.

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 $r \times c$ 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 *p* 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 $y_j \in \Gamma$ that is sampled, define the binary outcome $z_j = 1$ if $y_j \in \Gamma^*$; 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:

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

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

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

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

$$CI = \hat{p}_2 \pm 2.576\hat{\sigma}/(\sqrt{M}) \qquad \textbf{Equation 9.8}$$

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 $\hat{\sigma}$ 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 $\alpha\,\%$ confidence interval for the exact p value is

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

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

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

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 *p* value** is obtained as

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

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 $(r-1) \times (c-1)$; 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.

10 Unordered R x C Contingency Tables

The tests in this chapter are applicable to $r \times c$ 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 in this setting because they have good power against the omnibus alternative, or universal hypothesis, that the row and column classifications are not independent. Subsequent chapters deal with tests that have good power against more specific alternatives.

Available Tests

Exact Tests offers three tests for analyzing unordered $r \times c$ contingency tables. They are the Pearson chi-square test, the likelihood-ratio test, and Fisher's exact test. Asymptotically, all three tests follow the chi-square distribution with $(r-1)(c-1)$ degrees of freedom. Both 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. Table 10.1 shows the three available tests, the procedure from which they can be obtained, and a bibliographical reference for each test.

Table 10.1 Available tests

Test	Procedure	Reference
Pearson chi-square test	Crosstabs	Agresti (1990)
Likelihood-ratio test	Crosstabs	Agresti (1990)
Fisher's exact test	Crosstabs	Freeman and Halton (1951)

When to Use Each Test

Any of the three tests, Pearson, likelihood-ratio, or Fisher's, may be used when both the row and column classifications of the $r \times c$ contingency table are unordered. All 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:

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

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

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

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

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

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

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 jth category, $j = 1, 2, \ldots c$, 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 powerful tests, such as the Kruskal-Wallis test, are available if the c categories have a natural ordering (for example, levels of toxicity). Such tests are discussed in Chapter 11 and Chapter 12.

Oral Lesions Data

The exact, Monte Carlo, and asymptotic versions of the Pearson chi-square test, the likelihood-ratio test, and Fisher's exact test can be illustrated with the following sparse data set. Suppose that data were obtained on the location of oral lesions, in house-to-house surveys in three geographic regions of rural India. These data are displayed here in the form of a 9×3 contingency table, as shown in Figure 10.1. The variables shown in the table are *site,* which indicates the specific site of the oral lesion, and *region*, which indicates the geographic region. *Count* represents the number of patients with oral lesions at a specific site and living in a specific geographic region.

Figure 10.1 Crosstabulation of oral lesions data set

Site of Lesion * Geographic Region Crosstabulation

Count

		Geographic Region		
		Kerala	Gujarat	Andhra
Site of Lesion	Labial Mucosa		1	
	Buccal Mucosa	8	1	8
	Commissure		1	
	Gingiva		1	
	Hard Palate		1	
	Soft Palate		1	
	Tongue		1	
	Floor of Mouth	1		1
	Alveolar Ridge	1		1

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 9×3 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 p values.

Pearson Chi-Square Test

The Pearson chi-square test is perhaps the most commonly used procedure for testing null hypotheses of the form shown in Equation 10.1 or Equation 10.2 for independence of row and column classifications in an unordered $r \times c$ contingency table. For any observed $r \times c$ table, the test statistic, $D(x)$, is denoted as $CH(x)$ and is computed by the formula

$$CH(x) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - m_i n_j / N)^2}{m_i n_j / N} \qquad \textbf{Equation 10.3}$$

For the 9×3 contingency table of oral lesions data displayed in Figure 10.1, $CH(x) = 22.1$. The test statistic and its corresponding asymptotic and exact p values are shown in Figure 10.2.

Figure 10.2 Exact and asymptotic Pearson chi-square test for oral lesions data

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)
Pearson Chi-Square	22.099[1]	16	.140	.027

1. 25 cells (92.6%) have expected count less than 5. The minimum expected count is .26.

The results show that the observed value of the test statistic is $CH(x) = 22.1$. This statistic has an asymptotic chi-square distribution with 16 degrees of freedom.

The asymptotic p value is based on the 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 $CH(x) = 22.1$. The p value of 0.14 implies that there is no row-by-column interaction. However, this p value cannot be trusted because of the sparseness of the observed contingency table.

The exact p value is shown in the portion of the output entitled *Exact Sig. (2-tailed)*. It is defined by Equation 9.4 as the permutational probability $\Pr(CH(y) \geq 22.1 \,|\, y \in \Gamma)$. The

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 $r \times c$ contingency table. For any observed $r \times c$ contingency table, the test statistic, $D(x)$, is denoted as $LI(x)$ and is computed by the formula

$$LI(x) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log\left(\frac{x_{ij}}{m_i n_j / N}\right) \qquad \textbf{Equation 10.4}$$

For the oral lesions data displayed in Figure 10.1, $LI(x) = 23.3$. The test statistic and its corresponding asymptotic and exact p values are shown in Figure 10.4.

Figure 10.4 Results of likelihood-ratio test for oral lesions data

Chi-Square Tests

		Values			
		Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)
Statistics	Likelihood Ratio	23.297	16	.106	.036

The output shows that the observed value of the test statistic is $LI(x) = 23.3$. 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 $LI(x) = 23.3$. The p value of 0.106 implies that there is no row-by-column interaction. However, this p value cannot be trusted because of the sparseness of the observed contingency table.

The exact p value is defined by Equation 9.4 as the permutational probability $\Pr(LI(y) \geq 23.3 | y \in \Gamma)$. The exact p value is 0.036, showing that there is a significant interaction between the site of 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 10.5 shows an unbiased estimate of the exact p value for the likelihood-ratio test based on a crude Monte Carlo sample of 10,000 tables from the reference set.

Figure 10.5 Estimate of exact p value for likelihood-ratio test based on Monte Carlo sampling

Chi-Square Tests

		Values					
					Monte Carlo Significance (2-tailed)		
						99% Confidence Interval	
		Value	df	Asymp. Sig. (2-tailed)	Sig.	Lower Bound	Upper Bound
Statistics	Likelihood Ratio	23.297	16	.106	.035[2]	.030	.039

2. Based on 10000 and seed 2000000 ...

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 $r \times c$ 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 $r \times c$ 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 $r \times c$ contingency table, the test statistic, $D(x)$, is denoted as $Fl(x)$ and is computed by the formula

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

where

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

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 $\Pr(FI(y) \geq 19.72 | y \in \Gamma)$. The exact p value is 0.010, showing that there is a significant interaction between the site of the lesion and the geographic region. The asymptotic result was off the mark and failed to demonstrate a significant outcome. 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 10.7 shows an unbiased estimate of the exact p value for Fisher's exact test based on a crude Monte Carlo sample of 10,000 tables from the reference set.

Figure 10.7 Monte Carlo estimate of Fisher's exact test for oral lesions data

Chi-Square Tests

		Values			
			Monte Carlo Significance (2-tailed)		
				99% Confidence Interval	
		Value	Sig.	Lower Bound	Upper Bound
Statistics	Fisher's Exact Test	19.721	.010[1]	.007	.013

1. Based on 10000 and seed 2000000 ...

The Monte Carlo method produces a 99% 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 0.007 to 0.013. 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.

11 Singly Ordered R x C Contingency Tables

The test in this chapter is applicable to $r \times c$ 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 $r \times c$ contingency tables in which the rows (r) are unordered but the columns (c) have a natural ordering. Although the logic of the Kruskal-Wallis test can be applied to singly ordered contingency tables, this test is performed through the Nonparametric Tests: Tests for Several Independent Samples procedure. (See Siegal and Castellan, 1988.)

When to Use the Kruskal-Wallis Test

Use the Kruskal-Wallis test for an $r \times c$ contingency table in which the rows (r) are unordered but the columns (c) are ordered. Note that it is very important to keep the columns ordered, not the rows. In this chapter, the Kruskal-Wallis test is applied to ordinal categorical data. See Chapter 8 for a discussion of using this test for continuous data.

Statistical Methods

The data consist of c categorical responses generated by subjects in r populations, and cross-classified into an $r \times c$ contingency table, as shown in Table 9.1. The c categorical responses are usually ordered, whereas the r populations are not. Suppose there are m_i subjects in population i and each subject generates a multinomial response falling into one of c ordered categories with respective multinomial probabilities of $\Pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ic})$ for $i = 1, 2, \ldots, r$.

The null hypothesis is

$$H_0{:}\Pi_1 = \Pi_2 = \ldots = \Pi_r \qquad \textbf{Equation 11.1}$$

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 $i = 1, 2, \ldots, r$, let

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

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

$$H_1\text{: for at least one}(i_1, i_2)\text{ pair,}\Upsilon_{i_1 j} \geq \Upsilon_{i_2 j}, j = 1, 2, \ldots c \qquad \textbf{Equation 11.2}$$

with strict inequality for at least one j. In other words, you want to reject H_0 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.

Figure 11.1 Crosstabulation of tumor regression data

Chemotherapy Regimen * Tumor Regression Crosstabulation

Count

		Tumor Regression		
		No Response	Partial Response	Complete Response
Chemotherapy Regimen	CTMX	2		
	CCNU	1	1	
	MTX	3		
	CTX+CCNU	2	2	
	CTX+CCNU+MTX	1	1	4

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 Response*, 2=*Partial Response*, and 3=*Complete Response*. Note that although variable labels are displayed, these variables must be numeric.

Figure 11.2 Tumor regression data displayed in the Data Editor

	chemo	regressn
1	CTX	No Response
2	CTX	No Response
3	CCNU	No Response
4	CCNU	Partial Response
5	MTX	No Response
6	MTX	No Response
7	MTX	No Response
8	CTX+CCNU	No Response
9	CTX+CCNU	No Response
10	CTX+CCNU	Partial Response
11	CTX+CCNU	Partial Response
12	CTX+CCNU+MTX	No Response
13	CTX+CCNU+MTX	Partial Response
14	CTX+CCNU+MTX	Complete Response
15	CTX+CCNU+MTX	Complete Response
16	CTX+CCNU+MTX	Complete Response
17	CTX+CCNU+MTX	Complete Response

Small pilot studies like this one are frequently conducted as a preliminary step to planning a large-scale randomized clinical trial. The test in this section may be used to determine whether or not the five drug regimens are significantly different with respect to their tumor regression rates. Notice how appropriate the alternative hypothesis, shown in Equation 11.2, is for this situation. It can be used to detect departures from the null hypothesis in which one or more drugs shift the responses from no response to partial or complete responses. The results of the Kruskal-Wallis test are shown in Figure 11.3.

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 $\Pr(T \geq 8.682 | y \in \Gamma)$. 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 $\Pr(T \geq 8.682)$. 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.

Figure 11.4 Monte Carlo results for tumor regression data

Test Statistics[1,2]

	Chi-Square	df	Asymp. Sig.	Monte Carlo Sig.		
				Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound
Tumor Regression	8.682	4	.070	.043[3]	.037	.048

1. Kruskal Wallis Test

2. Grouping Variable: Chemotherapy Regimen

3. Based on 10000 sampled tables with starting seed 20000000.

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 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 0.037 to 0.048. More tables could be sampled from the reference set 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 and column interaction in this contingency table. The asymptotic inference produced a p value of 0.070, and thus failed to demonstrate a statistically significant row-by-column interaction.

12 Doubly Ordered R x C Contingency Tables

The tests in this chapter are applicable to $r \times c$ contingency tables in which both the rows and columns are ordered. A typical example would be an $r \times c$ 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 $r \times c$ 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 12.1Available 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 $r \times c$ 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 $r \times c$ 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 numerical example in "Linear-by-Linear Association Test" on p. 161.

Statistical Methods

Suppose that each response must fall into one of c ordinal categories according to a multinomial distribution. Let m_i responses from population i fall into the c ordinal categories with respective multinomial probabilities of

$$\Pi_i = (\pi_{i1}, \pi_{i2}, \pi..., \pi_{ic})$$

for $i = 1, 2, ..., r$. The null hypothesis is

$$H_0: \Pi_1 = \Pi_2 = ... = \Pi_r \qquad \textbf{Equation 12.1}$$

To specify the alternative hypothesis, define

$$\Upsilon_{ij} = \sum_{l=1}^{j} \pi_{il} \qquad \textbf{Equation 12.2}$$

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

$$H_1: \Upsilon_{1j} \leq \Upsilon_{2j} \leq ... \leq \Upsilon_{rj}$$ **Equation 12.3**

or

$$H'_1: \Upsilon_{1j} \geq \Upsilon_{2j} \geq ... \geq \Upsilon_{rj}$$ **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 H_0 in favor of H'_1.

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.

Figure 12.1 Crosstabulation of dose-response data

Drug Dose * TOXICITY Crosstabulation

Count

		TOXICITY			
		Mild	Moderate	Severe	Death
Drug Dose	100	100	1		
	200	18	1	1	
	300	50	1	1	1
	400	50	1	1	1

Notice that there is a natural ordering across both the rows and the columns of the above 4×4 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* is an ordered grouping variable that indicates dose level, and *toxicity* is an ordered categorical variable with four values, where 1=*Mild*, 2=*Moderate*, 3=*Severe, and* 4=*Death*. Note that although value labels are displayed, these variables must be numeric. This is a large data set, with 227 cases, and therefore Figure 12.2 shows only a small subset of these data in order to illustrate the necessary data structure for the Jonckheere-Terpstra test. The full data set was used in the following example.

Figure 12.2 Dose-response data, displayed in the Data Editor

	dose	toxicity
1	100 mg	Mild
2	100 mg	Mild
3	200 mg	Severe
4	100 mg	Mild
5	400 mg	Moderate
6	400 mg	Mild
7	100 mg	Mild
8	100 mg	Mild
9	300 mg	Mild
10	300 mg	Severe
11	200 mg	Mild
12	100 mg	Mild
13	100 mg	Mild
14	100 mg	Moderate
15	100 mg	Mild
16	400 mg	Mild
17	400 mg	Mild

You can run the Jonckheere-Terpstra test on the dose-response data shown in Figure 12.2. The results are shown in Figure 12.3.

Figure 12.3 Results of Jonckheere-Terpstra test for dose-response data

	Number of Levels in Drug Dose	N	Observed J-T Statistic	Mean J-T Statistic	Std. Deviation of J-T Statistic	Std. J-T Statistic	Asymp. Sig. (2-tailed)	Exact Significance (2-tailed)	Exact Sig. (1-tailed)	Point Probability
TOXICITY	4	227	9127.000	8827.500	181.760	1.648	.099	.100	.049	.000

1. Grouping Variable: Drug Dose

The value of the observed test statistic, defined by Equation 8.38, is $t = 9127$, the mean is $E(T) = 8828$, the standard deviation is 181.8, and the standardized test statistic, calculated by Equation 8.41, is $t^* = 1.65$. The standardized statistic is normally distributed with a mean of 0 and a variance of 1, while its square is chi-square distributed with one degree of freedom.

The asymptotic two-sided p values are evaluated as the tail areas under a standard normal distribution. In calculating the one-sided p value, which is not displayed in the output, a choice must be made as to whether to select the left tail or the right tail at the observed value $t^* = 1.65$. In Exact Tests, this decision is made by selecting the tail area with the smaller probability. Thus, the asymptotic one-sided p value is calculated as

$$\tilde{p}_1 = \min\{\Phi(t^*), 1 - \Phi(t^*)\} \qquad \textbf{Equation 12.5}$$

where $\Phi(z)$ is the tail area from $-\infty$ to z under a standard normal distribution. In the present example, it is the right tail area that is the smaller of the two, so that the asymptotic one-sided p value is evaluated as the normal approximation to $\Pr(T^* \geq 1.65)$, which works out to 0.0490. The asymptotic two-sided p value is defined as double the one-sided:

$$\tilde{p}_2 = 2\tilde{p}_1 = 0.0994 \qquad \textbf{Equation 12.6}$$

Since the square of a standard normal variate is a chi-square variate with one degree of freedom, an equivalent alternative way to compute the asymptotic two-sided p value is to evaluate the tail area to the right of $(1.65)^2$ from a chi-square distribution with one degree of freedom. It is easy to verify that this too will yield 0.099 as the asymptotic two-sided p value.

The exact one-sided p value is computed as the smaller of two permutational probabilities:

$$p_1 = \min\{\Pr(T^* \leq 1.65), \Pr(T^* \geq 1.65)\} \qquad \textbf{Equation 12.7}$$

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

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

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

Jonckheere-Terpstra Test[1]

	Number of Levels in Drug Dose	N	Observed J-T Statistic	Mean J-T Statistic	Std. Deviation of J-T Statistic	Std. J-T Statistic	Asymp. Sig. (2-tailed)	Monte Carlo Sig. (2-tailed)			Monte Carlo Sig. (1-tailed)		
									99% Confidence Interval			99% Confidence Interval	
								Sig.	Lower Bound	Upper Bound	Sig.	Lower Bound	Upper Bound
TOXICITY	4	227	9127.000	8827.500	181.760	1.648	.099	.101[2]	.093	.109	.051[2]	.045	.057

[1] Grouping Variable: Drug Dose
[2] Based on 10000 sampled tables with starting seed 2000000.

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

$$LL(x) = \sum_{i=1}^{r}\sum_{j=1}^{c} u_i v_j x_{ij}$$

Equation 12.9

where $u_i, i = 1, 2, \ldots, r$ are arbitrary row scores, and $v_j, j = 1, \ldots, c$ are arbitrary column scores. Under the null hypothesis of no row-by-column interaction, the linear-by-linear statistic has a mean of

$$E(LL(X)) = N^{-1}\left(\sum_{i=1}^{r} u_i m_i\right)\left(\sum_{j=1}^{c} v_j n_j\right)$$

Equation 12.10

and a variance of

$$\text{var}(LL(X)) = (N-1)^{-1}\left[\sum_i u_i^2 m_i - \frac{(\Sigma_i u_i m_i)^2}{N}\right]\left[\sum_j v_j^2 n_j - \frac{(\Sigma_j v_j n_j)^2}{N}\right]$$

Equation 12.11

See Agresti (1990) for more information. The asymptotic distribution of

$$LL^*(X) = \frac{LL(X) - E(LL(X))}{\sqrt{\text{var}(LL(X))}}$$

Equation 12.12

is normal, with a mean of 0 and a variance of 1, where *LL** denotes the standardized version of *LL*. The square of the normalized statistic is distributed as chi-square with one degree of freedom.

Next, run the linear-by-linear association test on the dose-response data shown in Figure 12.1. The results are shown in Figure 12.5.

Figure 12.5 Results of linear-by-linear association test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Linear-by-Linear Association	3.264[2]	1	.071	.079	.044	.012

2. Standardized stat. is 1.807 ...

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 $LL^* = 1.807$. This value is normally distributed with a mean of 0 and a variance of 1. The chi-square value, 3.264, is the square of this standardized value. The asymptotic two-sided p value is calculated under a chi-square distribution.

The exact one- and two-sided p values are also displayed in the output. The exact one-sided p value is computed as the smaller of two permutational probabilities:

$$p_1 = \min\{(\Pr|LL^*(y) \le 1.807|y \in \Gamma), \Pr(LL^*(y) \ge 1.807|y \in \Gamma)\} \quad \textbf{Equation 12.13}$$

In the present example, the smaller permutational probability is the one that evaluates the right tail. This value is 0.044. The exact one-sided p value is the point probability $\Pr(LL^*(X) = 1.807)$. This probability, 0.012, 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. (For more information on the role of the mid-p method in exact inference, see Lancaster, 1961; Pratt and Gibbons; 1981, and Miettinen, 1985.) In Equation 12.14, the point probability is the exact two-sided p value

$$p_2 = \Pr(|LL^*(X)| \ge 1.807) = 0.0792 \quad \textbf{Equation 12.14}$$

Notice that this definition will produce the same answer as Equation 9.4, with $D(y) = (LL^*(y))^2$ 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 12.6 displays an unbiased estimate of the exact one- and two-sided p values for the linear-by-linear association test based on a crude Monte Carlo sample of 10,000 tables from the reference set.

Figure 12.6 Monte Carlo results for linear-by-linear association test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Monte Carlo Significance (2-tailed)			Monte Carlo Significance (1-tailed)		
				Sig.	99% Confidence Interval		Sig.	99% Confidence Interval	
					Lower Bound	Upper Bound		Lower Bound	Upper Bound
Linear-by-Linear Association	3.264[3]	1	.071	.081[2]	.073	.088	.046[2]	.040	.051

2. Based on 10000 and seed 2000000 ...

3. Standardized stat. is 1.807 ...

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 value is contained in the interval 0.040 to 0.051. The Monte Carlo point estimate of the exact two-sided p value is 0.081, and the corresponding 99% confidence interval is 0.073 to 0.088. More tables could be sampled from the reference set to further narrow the widths of these intervals. One important advantage of the linear-by-linear association test over the Jonckheere-Terpstra test is its ability to specify arbitrary row and column scores. Suppose, for example, that you want to penalize the greater toxicity levels by greater amounts through the unequally spaced scores (1, 3, 9, 27). The crosstabulation of the new data is shown in Figure 12.7.

Figure 12.7 Drug dose data penalized at greater toxicity levels

Drug Dose * TOXICITY Crosstabulation

Count

		TOXICITY			
		Mild 1	Severe 3	9	27
Drug Dose	100 mg	100	1		
	200 mg	18	1	1	
	300 mg	50	1	1	1
	400 mg	50	1	1	1

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

Chi-Square Tests

	Value	df	Asymp. Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-tailed)	Point Probability
Linear-by-Linear Association	3.008[2]	1	.083	.078	.050	.005

2. Standardized stat. is 1.734 ...

Observe now that the one-sided asymptotic p value is 0.042, $(0.083)/2$, 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 ρ is a population parameter in a bivariate normal distribution, whereas Pearson's product moment coefficient R is an estimate of ρ, 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 $r \times c$ contingency table, as shown in Table 13.1.

Table 13.1 Observed r x c contingency table

Row Number	Column Number				Row Totals	Row Scores
	Col_1	**Col_2**	…	**Col_c**		
Row_1	x_{11}	x_{12}	…	x_{1c}	m_1	u_1
Row_2	x_{21}	x_{22}	…	x_{2c}	m_2	u_2
.	.	.	…	.	.	.
.	.	.		.	.	.
.	.	.		.	.	.
Row_r	x_{r1}	x_{r2}	…	x_{rc}	m_r	u_r
Col_Totals	n_1	n_2	…	n_c	N	
Col_Scores	v_1	v_2	…	v_c		

This $r \times c$ table is formed from N observations cross-classified into row categories (r) and column categories (c), with x_{ij} 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 nevertheless convenient to create artificial crosstabulations from them in order to present one unified method of defining and computing measures of association. To see this, let A, B represent a pair of random variables following a bivariate distribution, and let $\{(a_1,b_1), (a_2,b_2), \ldots (a_N,b_N)\}$ be N pairs of observations drawn from this bivariate distribution. The data may contain ties. Moreover, the original data might be replaced by rank scores. To accommodate these possibilities, let $(u_1 < u_2 < \ldots < u_r)$ be r distinct scores assumed by the A component of the data series, sorted in ascending order. The u_i's might represent the raw data, the data replaced by ranks, or the raw data replaced by arbitrary scores. When there are no ties, r will equal N. Similarly, let $(v_1 < v_2 < \ldots < v_c)$ be c distinct scores assumed by the B component of the data series. Now the bivariate data can be cross-classified into an $r \times c$ contingency table such as Table 13.1, with u_i as the score for row i and v_j as the score for column j.

For example, consider the bivariate data set shown in Figure 13.1. This data set is adapted from Siegel and Castellan (1988) with appropriate alterations to illustrate the effect of ties. The original data are shown in Chapter 14. Each subject was measured on two scales—authoritarianism and social status striving—and the goal was to estimate

the correlation between these two measures. Figure 13.1 shows the data displayed in the Data Editor. *Author* contains subjects' measurements on the authoritarianism scale, and *status* contains subjects' measurements on the social status striving scale. Figure 13.2 shows the same data set crosstabulated as a 5×5 contingency table.

Figure 13.1 Bivariate data set

subject	author	social
1	82	46
2	87	46
3	87	39
4	40	56
5	111	65
6	113	88
7	111	88
8	82	46

Figure 13.2 Crosstabulation of bivariate data set

Authoritarianism * social status striving Crosstabulation

Count

		social status striving				
		39	46	56	65	88
Authoritarianism	40			1		
	82		2			
	87	1	1			
	111				1	1
	113					1

The original data consist of $N = 8$ pairs of observations. These data are replaced by an equivalent contingency table. Because these data contain ties, the contingency table is 5×5 instead of 8×8. Had the data been free of ties, every row and column sum would have been unity, and the equivalent contingency table would have been 8×8. In this sense, the contingency table is not a natural representation of paired continuous data, since it can artificially expand N bivariate pairs into an $N \times N$ rectangular array. However, it is convenient to represent the data in this form, since it provides a consistent notation for defining all of the measures of association and related statistics that you will be estimating.

Point Estimates

Maximum-likelihood theory is used to estimate each measure of association. For this purpose, Table 13.1 is constructed by taking N samples from a multinomial distribution and observing counts x_{ij} in cells (i,j) with the probability π_{ij}, where $\Sigma_{i,j}\pi_{ij} = 1$. Measures of association are functions of these cell probabilities. A maximum-likelihood estimate (MLE) is provided for each measure, along with an asymptotic standard error (ASE1) evaluated at the MLE. All of the measures of association defined from ordinal data in Chapter 14 and all of the measures of agreement in Chapter 16 fall in the range of –1 to +1, with 0 implying that there is no association, –1 implying a perfect negative association, and +1 implying a perfect positive association.

All of the measures of association defined from nominal data in Chapter 15 fall in the range of 0 to 1, with 0 implying that there is no association and 1 implying perfect association.

Exact P Values

Exact p values are computed by the methods described in Chapter 9. First, the reference set, Γ, is defined to be all $r \times c$ tables with the same margins as the observed table, as shown in Equation 9.1. Under the null hypothesis that there is no association, each table $y \in \Gamma$ has the hypergeometric probability $P(y)$, given by Equation 9.2. Then each table $y \in \Gamma$ is assigned a value $M(y)$ corresponding to the measure of association being investigated.

Nominal Data

For measures of association on nominal data, only two-sided p values are defined. The exact two-sided p value is computed by Equation 9.4, with $M(y)$ substituted for $D(y)$. Thus,

$$p_2 = \sum_{M(y) \geq M(x)} P(y) = \Pr\{M(y) \geq M(x)\} \qquad \textbf{Equation 13.1}$$

Ordinal and Agreement Data

For measures of association based on ordinal data and for measures of agreement, only two-sided p values are defined. Now $M(y)$ is a univariate test statistic ranging between –1 and +1, with a mean of 0. A negative value for $M(y)$ implies a negative association between the row and column variables, while a positive value implies a positive associa-

tion. The exact two-sided p value is obtained by Equation 9.4, with $M^2(y)$ substituted for $D(y)$. Thus,

$$p_2 = \sum_{M^2(y) \geq M^2(x)} P(y) = \Pr\{M^2(y) \geq M^2(x)\} \qquad \textbf{Equation 13.2}$$

An equivalent definition of the two-sided p value is

$$p_2 = \sum_{|M(y)| \geq |M(x)|} P(y) = \Pr\{|M(y)| \geq |M(x)|\} \qquad \textbf{Equation 13.3}$$

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 $M(y)$. 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 $M(y)$.

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. 139. 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 Γ^* is defined by

$$\Gamma^* = \{y \in \Gamma : M(y) \geq M(x)\} \qquad \textbf{Equation 13.4}$$

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

$$\Gamma^* = \{y \in \Gamma : |M(y)| \leq |M(x)|\} \qquad \textbf{Equation 13.5}$$

Asymptotic P Values

For measures of association based on nominal data, only two-sided p values are defined. These p values are obtained as tail areas of the chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

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 14.1Available 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 $\{(a_1, b_1)(a_2, b_2)\ldots(a_N, b_N)\}$ and crosstabulate them into the $r \times c$ contingency table displayed as Table 13.1, in which the u_i's are the distinct values assumed by A and the v_j'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, ρ, between A and B. This parameter is estimated by Pearson's product-moment correlation coefficient, shown in Equation 14.1. In this equation, m_i represents the marginal row total and n_j represents the marginal column total.

$$R = \frac{\Sigma_{i=1}^{r}\Sigma_{j=1}^{c}x_{ij}(u_i-\bar{u})(v_j-\bar{v})}{\sqrt{\Sigma_{i=1}^{r}m_i(u_i-\bar{u})^2}\sqrt{\Sigma_{j=1}^{c}n_j(v_j-\bar{v})^2}} \qquad \textbf{Equation 14.1}$$

where

$$\bar{u} = \sum_{i=1}^{r} m_i u_i / N \text{ and } \bar{v} = \sum_{j=1}^{c} n_j v_j / N \qquad \textbf{Equation 14.2}$$

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 Algorithms on the Help menu.

You now compute Pearson's product-moment correlation coefficient for the first seven pairs of observations of the authoritarianism and social status striving data discussed in Siegel and Castellan (1988). The data are shown in Figure 14.1. *Author* contains subjects' scores on the authoritarianism scale, and *social* contains subjects' scores on the social status striving scale.

Figure 14.1 Subset of social status striving data

subject	author	social
1	82	42
2	98	46
3	87	39
4	40	37
5	116	65
6	113	88
7	111	86

The results are shown in Figure 14.2

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

subject	author	social
1	82	42
2	98	46
3	87	39
4	40	37
5	116	65
6	113	88
7	111	86
8	83	56
9	85	62
10	126	92
11	106	54
12	117	81

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 corresponding exact output is shown in Figure 14.5.

Figure 14.4 Correlations for complete social status striving data using the Monte Carlo method

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Monte Carlo Sig.		
						Sig.	99% Confidence Interval	
							Lower Bound	Upper Bound
Interval by Interval	Pearson's R	.775	.060	3.872	.003[1]	.002[2]	.001	.003
N of Valid Cases		12						

1. Based on normal appoximation.
2. Based on 10000 sampled tables with starting seed of 2000000.

Figure 14.5 Exact results for correlations for complete social status striving data

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Exact Sig.
Interval by Interval	Pearson's R	.775	.060	3.872	.003[1]	.001
N of Valid Cases		12				

1. Based on normal approximation.

Spearman's Rank-Order Correlation Coefficient

If you are reluctant to make the assumption of bivariate normality, you can use Spearman's rank-order correlation coefficient instead of Pearson's product-moment correlation coefficient. The only difference between the two measures of association is that Pearson's measure uses the raw data, whereas Spearman's uses ranks derived from the raw data. Specifically, if the data are represented in the crosstabular form of Table 13.1, Pearson's measure uses the raw data as the u_i and v_j scores, while Spearman's measure uses

$$u_i = m_1 + m_2 + \ldots + m_{i-1} + (m_i + 1)/2 \qquad \textbf{Equation 14.3}$$

for $i = 1, 2, \ldots r$, and

$$v_j = n_1 + n_2 + \ldots + n_{j-1} + (n_j + 1)/2 \qquad \textbf{Equation 14.4}$$

for $j = 1, 2, \ldots c$. 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 data with their ranks. Variable *rauthor* contains the ranks for *author*, the authoritarianism scores, and variable *rsocial* contains the ranks for *social*, the social status striving scores.

Figure 14.6 Raw data and rank scores for eight-case subset of social status striving data

	subject	author	rauthor	social	rsocial
1	1	82	2.5	46	3.0
2	2	87	4.5	46	3.0
3	3	87	4.5	39	1.0
4	4	40	1.0	56	5.0
5	5	111	6.5	65	6.0
6	6	113	8.0	88	7.5
7	7	111	6.5	88	7.5
8	8	82	2.5	46	3.0

Notice that tied ranks have been replaced by mid-ranks. These same rank scores could be obtained by crosstabulating *author* with *social*, and applying Equation 14.3 and Equation 14.4. The crosstabulation of the rank scores is shown in Figure 14.7.

Figure 14.7 Crosstabulation of rank scores for eight-case subset of social status striving data

RANK of AUTHOR * RANK of SOCIAL Crosstabulation

Count

		RANK of SOCIAL				
		1.0	3.0	5.0	6.0	7.5
RANK of AUTHOR	1.0			1		
	2.5		2			
	4.5	1	1			
	6.5				1	1
	8.0					1

Figure 14.8 shows the point and interval estimates for Spearman's correlation coefficient for these data. The exact and asymptotic p values for testing the null hypothesis that there is no correlation are also shown.

Figure 14.8 Exact results for Spearman's correlation coefficient for eight-case subset of social status striving data

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Exact Sig.
Ordinal by Ordinal	Spearman Correlation	.594	.309	1.808	.121[1]	.125
N of Valid Cases		8				

1. Based on normal approximation.

The Spearman rank-order correlation coefficient has a point estimate of $R = 0.594$. The exact two-sided p value is evaluated by Equation 9.4, as discussed in "Exact P Values" on p. 172 in Chapter 13. Its value is 0.125 and indicates that the correlation coefficient is not significantly different from 0. The corresponding asymptotic two-sided p value was 0.121.

As the number of paired observations grows, it becomes increasingly difficult to compute exact p values (i, j), and the Monte Carlo option is a better choice. Figure 14.9 shows the Monte Carlo results for the larger data set of 12 pairs of observations in Figure 14.3. The Monte Carlo sample size was 10,000. There is practically no difference between the Monte Carlo and exact p values.

Figure 14.9 Monte Carlo results for Spearman's correlation coefficient for complete social status striving data

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Monte Carlo Sig.		
						Sig.	99% Confidence Interval	
							Lower Bound	Upper Bound
Ordinal by Ordinal	Spearman Correlation	.818	.092	4.500	.001[1]	.002[2]	.001	.003
N of Valid Cases		12						

1. Based on normal approximation.

2. Based on 100000 and seed 2000000.

Kendall's W

All of the measures of association in this chapter are formed from a sequence of paired observations. Sometimes, however, the data consist of $K > 2$ related samples rather than just two related samples. Kendall's *W*, also known as Kendall's coefficient of concordance, is a measure of association specially developed for this situation. It bears a close relationship to Spearman's rank-order correlation coefficient. For $K > 2$ related samples of data, you could form $K!/2!(K-2)!$ distinct pairs of samples, and each pair would yield a value for Spearman's rank-order correlation coefficient. Let $\text{ave}(R_S)$ denote the average of all these Spearman correlation coefficients. Then you can show that, if there are no ties in the data,

$$\text{ave}(R_S) = \frac{KW - 1}{K - 1} \qquad \textbf{Equation 14.5}$$

Kendall's *W* is discussed in greater detail in Chapter 7, in the section "Kendall's W" on p. 106, where a numerical example is also provided.

Kendall's Tau and Somers' d Coefficients

Kendall's tau and Somers' *d* coefficients are alternatives to Pearson's product-moment correlation coefficient and Spearman's rank-order correlation coefficient for ordinal data. The main distinction between these measures and Pearson's or Spearman's measures is that you can compute the former without specifying numerical values for the row scores, u_i, or the column scores, v_j. All that is needed is an implicit ordering of the data. On the other hand, Equation 14.1, Equation 14.3, and Equation 14.4 relate the row and column scores explicitly to the computation of Pearson's and Spearman's coefficients.

Suppose that you have observed the $r \times c$ contingency table displayed as Table 9.1. Kendall's tau and Somers' *d* are both based on the difference between concordant and discordant pairs of observations in this contingency table. Since the rows and columns of the contingency table are ordered, the location of any cell (*h,k*) relative to any other cell (*i,j*) determines whether the observations in the two cells are concordant or discordant. For example, if $h < i$ and $k < j$, both members of a paired observation falling in cell (*h,k*) are smaller than the corresponding members of the paired observation falling in cell (*i,j*). Thus, the two pairs are concordant. On the other hand, if $h < i$ and $k > j$, the first member of the (*h,k*) pair is smaller, while the second member is larger than corresponding members of the (*i,j*) pair. The formula

$$C_{ij} = \sum_{h<i}\sum_{k<j} x_{hk} + \sum_{h>i}\sum_{k>j} x_{hj} \qquad \textbf{Equation 14.6}$$

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

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

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

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

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

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

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, τ, τ_b, and τ_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 τ_b and τ_c variants were developed to correct for ties and for categorical data.

Kendall's τ_b coefficient is estimated by

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

where

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

and

$$D_c = N^2 - \sum_{j=1}^{c} n_j \qquad \textbf{Equation 14.12}$$

Kendall's τ_c coefficient is estimated by

$$T_c = \frac{q(P-Q)}{N^2(q-1)} \qquad \textbf{Equation 14.13}$$

where $q = \min(r, c)$.

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 asymmetry. Somers' *d* has three variants; one with the row variable *U* as the independent variable, one with the column variable *V* as the independent variable, and a symmetric version. The row-independent version of Somers' *d* is

$$D_{V/U} = \frac{P-Q}{D_r} \qquad \textbf{Equation 14.14}$$

The column-independent version of Somers' *d* is

$$D_{U/V} = \frac{P-Q}{D_c} \qquad \textbf{Equation 14.15}$$

The symmetric version of Somers' *d* is

$$D = \frac{P-Q}{(.5)(D_r + D_c)} \qquad \textbf{Equation 14.16}$$

Example: Smoking Habit Data

Observe that all variants of Kendall's tau and Somers' d are functions of $P - Q$. 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. Consider the crosstabulation shown in Figure 14.10 for the status of the smoking habit and the length of the smoking habit. This data set was extracted from Siegel and Castellan (1988). For convenience, only 96 subjects with a smoking habit between 10 and 25 years in duration have been considered. The variables in the table are ***status***, which indicates the status of the smoking habit, with three categories (*successful quitter*, *in-process quitter*, and *unsuccessful quitter*), and ***years,*** which indicates the duration of the smoking habit.

Figure 14.10 Crosstabulation of cessation and years of smoking for subset of data

Count

		Years of Smoking Habit		
		10 to 14	15 to 19	20 to 25
Status of Smoking Habit	Successful Quitter	22	9	8
	In-process Quitter	2	1	3
	Unsuccessful Quitter	14	21	16

Figure 14.11 shows the results for the Kendall's tau-*b*, Kendall's tau-*c*, and all three variants of Somers' d for these data. The exact and asymptotic p values for testing the null hypothesis that there is no correlation are also shown.

Figure 14.11 Kendall's tau and Somers' d for subset of smoking data

Directional Measures

			Value	Asymp. Std. Error[1]	Approx. T[2]	Approx. Sig.	Exact Sig.
Ordinal by Ordinal	Somers' d	Symmetric	.214	.091	2.372	.018	.023
		Status of Smoking Habit Dependent	.196	.083	2.372	.018	.023
		Years of Smoking Habit Dependent	.236	.100	2.372	.018	.023

1. Not assuming the null hypothesis.
2. Using the asymptotic standard error assuming the null hypothesis.

Figure (Continued)

Symmetric Measures

		Value	Asymp. Std. Error[1]	Approx. T[2]	Approx. Sig.	Exact Sig.
Ordinal by Ordinal	Kendall's tau-b	.215	.091	2.372	.018	.023
	Kendall's tau-c	.194	.082	2.372	.018	.023
N of Valid Cases		96				

1. Not assuming the null hypothesis.
2. Using the asymptotic standard error assuming the null hypothesis.

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 values, and the Monte Carlo option is a better choice. Figure 14.12 shows the data for all 240 subjects who participated in the cessation of smoking study (Siegel and Castellan, 1988).

Figure 14.12 Full data set for cessation and years of smoking

Status of Smoking Habit * Years of Smoking Habit Crosstabulation

Count

		Years of Smoking Habit						
Status of Smoking Habit		1	2-4	5-9	10-14	15-19	20-25	> 25
Status of Smoking Habit	Successful Quitter	13	29	26	22	9	8	8
	In-Process Quitter	5	2	6	2	1	3	0
	Unsuccessful Quitter	1	9	16	14	21	16	29
Total		19	40	48	38	31	27	37

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

			Value	Asymp. Std. Error[1]	Approx. T[2]	Approx. Sig.	Monte Carlo Sig.		
							Sig.	99% Confidence Interval	
								Lower Bound	Upper Bound
Ordinal by Ordinal	Somers' d	Symmetric	.338	.046	7.339	.000	.000[3]	.000	.000
		Status of Smoking Habit Dependent	.282	.038	7.339	.000	.000[3]	.000	.000
		Years of Smoking Habit Dependent	.420	.058	7.339	.000	.000[3]	.000	.000

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.

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	Kendall's tau-b	.344	.047	7.339	.000	.000[3]	.000	.000
	Kendall's tau-c	.359	.049	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 0.0003 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

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

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.

15 Measures of Association for Nominal Data

Measures of association for nominal data are defined on $r \times c$ contingency tables like Table 13.1. However, these measures do not depend on the particular order in which the rows and columns are arranged, nor do they depend on row and column scores. Interchanging two rows or two columns does not alter these measures of association. Exact Tests provides the following measures of association between pairs of nominal categorical variables:

Contingency Coefficients. These coefficients are derived from the Pearson chi-square statistic. They include the Pearson coefficient, Cramér's V coefficient, and the phi coefficient.

Proportional Reduction in Prediction Error. Goodman and Kruskal's tau and the uncertainty coefficient are measures for assessing the power of one variable to predict the classification of members of the population with respect to a second variable.

These measures of association range between 0 and 1, with 0 signifying no association and 1 signifying perfect association.

Available Measures

Table 15.1 shows the available tests, the procedure from which they can be obtained, and a bibliographical reference for each test.

Table 15.1 Available tests

Measure of Association	Procedure	Reference
Contingency coefficients	Crosstabs	Liebetrau (1983)
Goodman and Kruskal's tau	Crosstabs	Bishop et al. (1975)
Uncertainty coefficient	Crosstabs	IMSL (1994)

Contingency Coefficients

All of the measures of association in this family are functions of the Pearson chi-square statistic $CH(x)$, specified by Equation 10.3. They include the phi contingency coefficient, the Pearson contingency coefficient, and Cramér's V contingency coefficient. All

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 $CH(y)$. 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 $CH(x)$. 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

$$\phi = \sqrt{\frac{CH(x)}{N}}$$ **Equation 15.1**

The minimum value assumed by ϕ 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 $r \times c$ tables. For the special case of the 2×2 table, Gibbons (1985) shows that ϕ is identical to the absolute value of Kendall's τ_b coefficient and is evaluated by the formula

$$\phi = \frac{x_{11}x_{22} - x_{12}x_{21}}{\sqrt{m_1 m_2 n_1 n_2}}$$ **Equation 15.2**

Notice from Equation 15.2 that, for the 2×2 contingency table, ϕ could be either positive or negative, which implies a positive or negative association in the 2×2 table.

The Pearson contingency coefficient is given by the formula

$$CC = \sqrt{\frac{CH(x)}{CH(x) + N}}$$ **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 $\sqrt{(q-1)/q}$, where $q = \min(r, c)$. Thus, the range of this contingency coefficient still depends on the dimensions of the $r \times c$ table. Cramér's V coefficient ranges between 0 and 1, with 0 signifying no association and 1 signifying perfect association. It is given by

$$V = \sqrt{\frac{CH(x)}{N(q-1)}}$$ **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 on the Help menu.

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

Survey Disposition * Finance Organization Crosstabulation

Count

		Finance Organization					
		AAA	AICPA	FAF	FASB	FEI	NAA
Survey Disposition	Completed	8	8	3	11	17	2
	Declined	2	5	1	2		13
	No Response	12	8	15	19	18	

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.

Figure 15.2 Phi and Cramér's V for first three columns for survey and finance organization data

Symmetric Measures

		Value	Approx. Sig.	Exact Significance
Nominal by Nominal	Phi	.359	.092	.090
	Cramer's V	.254	.092	.090
N of Valid Cases		62		

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

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

This coefficient ranges between 0 and 1, with 0 implying no reduction in row variance when the column category is known, and 1 implying complete reduction in row variance when the column category is known. An asymptotic confidence interval for the Goodman and Kruskal's tau can be obtained by computing the asymptotic standard error ASE1 and applying it to Equation 13.1. The exact two-sided p values for testing the null hypothesis that there is no association is obtained by substituting $\hat{\tau}_{R|C}(x)$ for $M(x)$ in Equation 13.1. The corresponding asymptotic two-sided p value is obtained by using the fact that $\hat{\tau}_{R|C}(x)$ converges to a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

Uncertainty Coefficient

The uncertainty coefficient is derived from the likelihood-ratio statistic and is an alternative way to measure the proportion of the variation in the row variable attributable to the column variable. It is estimated by

$$U_{R|C}(x) = \frac{\Sigma_{i=1}^{r}\Sigma_{i=1}^{c} x_{ij}\log(m_i n_j / N x_{ij})}{\Sigma_{i=1}^{r} m_i \log(m_i / N)} \qquad \textbf{Equation 15.6}$$

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

An asymptotic confidence interval for the uncertainty coefficient can be obtained by computing the asymptotic standard error ASE1 and applying it to Equation 13.1. The exact two-sided p values for testing the null hypothesis that there is no association is obtained by substituting $U_{R|C}(x)$ for $M(x)$ in Equation 13.1. The corresponding asymptotic two-sided p value is obtained by using the fact that $U_{R|C}(x)$ converges to a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

Example: Party Preference Data

The data set shown in Figure 15.4 illustrates the use of Goodman and Kruskal's tau and the uncertainty coefficient. The data set compares party preference with preferred cold war ally in Great Britain. These data are taken from Bishop, Fienberg, and Holland (1975).

Figure 15.4 Crosstabulation of party preference with preferred cold war ally

Count

		Preferred Cold War Ally	
		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.

Figure 15.5 Goodman and Kruskal's tau for party preference and preferred cold war ally data

Directional Measures

			Value	Asymp. Std. Error[1]	Approx. Sig.	Exact Significance
Nominal by Nominal	Goodman and Kruskal tau	Party Preference Dependent	.010	.006	.008[4]	.015
		Preferred Cold War Ally Dependent	.013	.010	.036[4]	.045

1. Not assuming the null hypothesis
4. Based on the chi-square approximation

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

Directional Measures

			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

1. Not assuming the null hypothesis
2. Using the asymptotic standard error assuming the null hypothesis
3. Likelihood ratio chi-square probability

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.

16 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 $r \times r$ 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 $r \times r$ contingency table. It is estimated by

$$K = \frac{N\Sigma_{i=1}^{r} x_{ii} - \Sigma_{i=1}^{r} m_i n_i}{N^2 - \Sigma_{i=1}^{r} m_i n_i} \qquad \textbf{Equation 16.1}$$

Notice that the kappa statistic does not depend on the off-diagonal elements of the observed contingency table. If the row classification is by one observer, and the column classification is by a second observer, this measure of agreement is determined entirely by the diagonal elements.

Example: Student Teacher Ratings

Consider the following data on student teachers who were rated by their supervisors, represented by variables *super1* and *super2*. The students were rated as *authoritarian*, *democratic*, or *permissive*. The full data set of 72 student teachers is available in Bishop, Fienberg, and Holland (1975). In the following example, a subset of 10 students is considered. The crosstabulated data are shown in Figure 16.1.

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

Rating by Supervisor 1 * Rating by Supervisor 2 Crosstabulation

Count

		Rating by Supervisor 2		
		Authoritarian	Democratic	Permissive
Rating by Supervisor 1	Authoritarian	3		1
	Democratic		2	
	Permissive	2		2

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

Figure 16.2 Kappa for student teacher ratings data

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Exact Significance
Measure of Agreement	Kappa	.531	.237	2.348	.019	.048
N of Valid Cases		10				

The value of kappa is estimated at $K = 0.531$. 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.

Figure 16.3 Crosstabulation of student teachers rated by supervisors (full data)

Rating by Supervisor 1 * Rating by Supervisor 2 Crosstabulation

Count

		Rating by Supervisor 2		
		Authoritarian	Democratic	Permissive
Rating by Supervisor 1	Authoritarian	17	4	8
	Democratic	5	12	
	Permissive	10	3	13

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

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.	Monte Carlo Significance		
						Sig.	99% Confidence Interval	
							Lower Bound	Upper Bound
Measure of Agreement	Kappa	.362	.091	4.329	.000	.0000[1]	.0000	.0005
N of Valid Cases		72						

1. Based on 10000 and seed 2000000 ...

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

CROSSTABS

Exact Tests Syntax

The /METHOD subcommand allows you to specify the method used to calculate significance levels. See the *Syntax Reference Guide* for a description of the full CROSSTABS syntax.

METHOD Subcommand

Displays additional results for each statistic requested. If no METHOD subcommand is specified, the standard asymptotic results are displayed. If fractional weights have been specified, results for all methods will be calculated on the weight rounded to the nearest integer.

MC Displays an unbiased point estimate and confidence interval based on the Monte Carlo sampling method, for all statistics. Asymptotic results are also displayed. When exact results can be calculated, they will be provided instead of the Monte Carlo results. See Appendix A for details of the situations under which exact results are provided instead of Monte Carlo results. Two optional keywords, CIN and SAMPLES, are provided if you choose /METHOD=MC.

CIN(*n*) Controls the confidence level for the Monte Carlo estimate. CIN is available only when /METHOD=MC is specified. CIN has a default value of 99.0. You can specify a confidence interval between 0.01 and 99.9, inclusive.

SAMPLES Specifies the number of tables sampled from the reference set when calculating the Monte Carlo estimate of the exact *p* value. Larger sample sizes lead to narrower confidence limits, but also take longer to calculate. You can specify any integer between 1 and 1,000,000,000 as the sample size. SAMPLES has a default value of 10,000.

EXACT Computes the exact significance level for all statistics, in addition to the asymptotic results. If both the EXACT and MC keywords are specified, only exact results are provided. Calculating the exact *p* value can be memory-intensive. If you have specified /METHOD=EXACT and find that you have insufficient memory to calculate results, you should first close any other applications that are currently running in order to make more memory available. You can also enlarge the size of your swap file (see your Windows manual for more information). If you still cannot obtain exact results, specify /METHOD=MC to obtain the Monte Carlo estimate of the exact *p* value. An optional TIMER keyword is available if you choose /METHOD=EXACT.

TIMER(n) Specifies the maximum number of minutes allowed to run the exact analysis for each statistic. If the time limit is reached, the test is terminated, no exact results are provided, and the application begins to calculate the next test in the analysis. TIMER is available only when /METHOD=EXACT is specified. You can specify any integer value for TIMER. Specifying a value of 0 for TIMER turns the timer off completely. TIMER has a default value of 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.

NPAR TESTS

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 NPAR TESTS syntax.

METHOD Subcommand

Displays additional results for each statistic requested. If no METHOD subcommand is specified, the standard asymptotic results are displayed.

MC Displays an unbiased point estimate and confidence interval based on the Monte Carlo sampling method, for all statistics. Asymptotic results are also displayed. When exact results can be calculated, they will be provided instead of the Monte Carlo results. See Appendix A for details of the situations under which exact results are provided instead of Monte Carlo results. Two optional keywords, CIN and SAMPLES, are provided if you choose `/METHOD=MC`.

CIN(*n*) Controls the confidence level for the Monte Carlo estimate. CIN is available only when `/METHOD=MC` is specified. You can specify a confidence interval between 0.01 and 99.9, inclusive.

SAMPLES Specifies the number of tables sampled from the reference set when calculating the Monte Carlo estimate of the exact *p* value. Larger sample sizes lead to narrower confidence limits, but also take longer to calculate. You can specify any integer between 1 and 1,000,000,000 as the sample size. SAMPLES has a default value of 10,000.

EXACT Computes the exact significance level for all statistics, in addition to the asymptotic results. If both the EXACT and MC keywords are specified, only exact results are provided. Calculating the exact *p* value can be memory-intensive. If you have specified `/METHOD=EXACT` and find that you have insufficient memory to calculate results, you should first close any other applications that are currently running in order to make more memory available. You can also enlarge the size of your swap file (see your Windows manual for more information). If you still cannot obtain exact results, specify `/METHOD=MC` to obtain the Monte Carlo estimate of the exact *p* value. An optional TIMER keyword is available if you choose `/METHOD=EXACT`.

TIMER(n) Specifies the maximum number of minutes allowed to run the exact analysis for each statistic. If the time limit is reached, the test is terminated, no exact results are provided, and the application begins to calculate the next test in the analysis. TIMER is available only when `/METHOD=EXACT` is specified. You can specify any integer value for TIMER. Specifying a value

of 0 for TIMER turns the timer off completely. TIMER has a default value of 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 subcommand is specified, one-tailed and two-tailed exact or Monte Carlo probabilities.

Syntax

- The minimum specification is a list of two variables. Variables must be polychotomous and must have more than two values. If the variables contain more than two values, the McNemar test is performed.
- If keyword WITH is not specified, each variable is paired with every other variable in the list.
- If WITH is specified, each variable before WITH is paired with each variable after WITH. If PAIRED is also specified, the first variable before WITH is paired with the first variable after WITH, the second variable before WITH with the second variable after WITH, and so on. PAIRED cannot be specified without WITH.
- With PAIRED, the number of variables specified before and after WITH must be the same. PAIRED must be specified in parentheses after the second variable list.

Operations

- The data consist of paired, dependent responses from two populations. The marginal homogeneity test tests the equality of two multinomial $c \times 1$ tables, and the data can be arranged in the form of a square $c \times c$ contingency table. A $2 \times c$ table is constructed for each off-diagonal cell count. The marginal homogeneity test statistic is computed for cases with different values for the two variables. Only combinations for which the values for the two variables are different are considered. The first row of each $2 \times c$ table specifies the category chosen by population 1, and the second row specifies the category chosen by population 2. The test statistic is calculated by summing the first row scores across all $2 \times c$ tables.

Example

```
NPAR TESTS /MH=V1 V2 V3
  /METHOD=MC.
```

- 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 A
Conditions for Exact Tests

There are certain conditions under which exact results are always provided, even when you have specified the Monte Carlo method either through the dialog box or through syntax. Table A.1 displays the conditions for the relevant tests under which exact results are always provided and a request for the Monte Carlo method is ignored.

Table A.1 Conditions under which exact tests are always provided

Test	Procedure	Condition
Binomial test	Nonparametric tests: Binomial Tests	Exact results are always provided
Fisher's exact test	Crosstabs	2×2 table
Likelihood-ratio test	Crosstabs	2×2 table
Linear-by-linear association test	Crosstabs	2×2 table
McNemar test	Nonparametric tests: Tests for two related samples	Exact results are always provided
Median test	Nonparametric tests: Tests for several related samples	$k = 2$ and $n \leq 30$
Pearson chi-square test	Crosstabs	2×2 table
Sign test	Nonparametric tests: Tests for two related samples	$n \leq 25$
Wald-Wolfowitz runs test	Nonparametric tests: Tests for two independent samples	$n \leq 30$

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 network algorithm for the exact treatment of the $2 \times k$ contingency table. *Communications in Statistics*, 9:6, 649–664.

Mehta, C. R., and N. R. Patel. 1983. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association*, 78:382, 427–434.

Mehta, C. R., N. R. Patel, and A. Tsiatis. 1984. Exact significance testing to establish treatment equivalence ordered categorical data. *Biometrics*, 40: 819–825.

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 $r \times c$ 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 $r \times c$ 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 $r \times c$ 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*.

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. *Computational Statistics and Data Analysis*, 5: 9–21.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

Breslow, N. E., and N. E. Day. 1980. The analysis of case-control studies. *IARC Scientific Publications*, No. 32. Lyon, France.

Chapman, J. W. 1976. A comparison of the chi-square, –2 log R, and multinomial probability criteria for significance tests when expected frequencies are small. *Journal of the American Statistical Association*, 71: 854–863.

Chernoff, H., and I. R. Savage. 1958. Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics*, 29: 972–994.

Cochran, W. G. 1936. The chi-square distribution for the binomial and Poisson series, with small expectations. *Annals of Eugenics, London*, 7: 207–217.

_____. 1954. Some methods for strengthening the common chi-square tests. *Biometrics*, 10: 417–454.

Conover, W. J. 1980. *Practical nonparametric statistics*. 2nd ed. New York: John Wiley and Sons.

Edgington, E. S. 1987. *Randomization tests*. 2nd ed. New York: Marcel Dekker.

Feynman, R. 1988. *What Do You Care What Other People Think?* New York: W. W. Norton and Co.

Fisher, R. A. 1924. The condition under which chi-square measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society*, 87: 442–450.

_____. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

_____. 1935a. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98: 39–54.

_____. 1935b. *The design of experiments*. Edinburgh: Oliver and Boyd.

_____. 1973. *Statistical methods and scientific inference*. 3rd ed. London: Collier Macmillan Publishers.

Freeman, G. H, and J. H. Halton. 1951. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, 38: 141–149.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32: 675–701.

Gastwirth, J. L. 1991. Statistical reasoning in a legal setting. *American Statistician*, February.

Gibbons, J. D. 1985. *Nonparametric statistical inference*. 2nd ed. New York: Marcel Dekker.

Good, P. 1993. *Permutation tests*. New York: Springer-Verlag.

Goodman, L. A. 1954. Kolmogorov-Smirnov tests for psychological research. *Psychological Bulletin*, 51: 160–168.

_____. 1968. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Association*, 63: 1091–1113.

Goodman, L. A., and W. H. Kruskal. 1979. *Measures of association for cross-classifications*. New York: Springer-Verlag.

Graubard, B. I., and E. L. Korn. 1987. Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics*, 43: 471–476.

Hajek, J. 1969. *Nonparametric statistics*. San Francisco: Holden-Day.

Hajek, J., and Z. Sidak. 1967. *Theory of rank tests*. New York: Academic Press, Inc.

Hollander, M., and D. A. Wolfe. 1973. *Nonparametric statistical methods*. New York: John Wiley and Sons.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30: 81–93.

Kendall, M. G., and B. Babington-Smith. 1939. The problem of *m* rankings. *Annals of Mathematical Statistics*, 10: 275–287.

Kendall, M. G., and A. Stuart. 1979. *The advanced theory of statistics*. 4th ed. New York: Macmillan Publishing Co. Inc.

Kruskal, W. H., and W. A. Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47: 583–621.

Kuritz, S. J., J. R. Landis, and G. G. Koch. 1988. A general overview of Mantel-Haenszel methods: Applications and recent developments. *Annual Review of Public Health*, 9: 123–60.

Lancaster, H. O. 1961. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56: 223–234.

Lehmann, E. L. 1975. *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day.

Liebetrau, A. M. 1983. *Measures of association*. Beverly Hills, Calif.: Sage Publications.

Little, R. J. A. 1989. Testing the equality of two independent binomial proportions. *The American Statistician*, 43: 283–288.

Makuch, R. W., and W. P. Parks. 1988. Response of serum antigen level to AZT for the treatment of AIDS. *AIDS Research and Human Retroviruses*, 4: 305–316.

Manley, B. F. J. 1991. *Randomization and Monte Carlo methods in biology*. London: Chapman and Hall.

Mehta, C. R., and N. R. Patel. 1983. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association*, 78:382, 427–434.

_____. 1986a. A hybrid algorithm for Fisher's exact test on unordered $r \times c$ contingency tables. *Communications in Statistics*, 15:2, 387–403.

_____. 1986b. FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Transactions on Mathematical Software*, 12:2, 154–161.

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.

Miettinen, O. S. 1985. *Theoretical epidemiology: Principles of occurrence research in medicine*. John Wiley and Sons, New York.

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 of Mathematical Statistics*, 11: 147–162.

Westfall, P. H., and S. S. Young. 1993. *Resampling-based multiple testing: Examples and methods for p value adjustment*. New York: John Wiley and Sons.

White, A. A., R. J. Landis, and M. M. Cooper. 1982. A note on the equivalence of several marginal homogeneity test criteria for categorical data. *International Statistical Review*, 50: 27–34.

Yates, F. 1984. Test of significance for 2×2 contingency tables. *Journal of the Royal Statistical Society, Series A*, 147: 426–463.

Yule, G. U. 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, Series A*, 75: 579.

Index

asymptotic method, 1
asymptotic one-sided *p* value
  K independent samples, 122, 129, 131
asymptotic one-sided *p* value
  Jonckheere-Terpstra test, 159
  Mann-Whitney test, 84
asymptotic *p* value, 12
  assumptions, 12
  defined, 16
  measures of association, 169
  obtaining, 7
  Pearson's chi-square, 16
  when to use, 16, 29–37
asymptotic two-sided *p* value
  K independent samples, 122
asymptotic two-sided *p* value
  Jonckheere-Terpstra test, 159
  K related samples, 101
  Mann-Whitney test, 84
  McNemar test, 69
  r x c tables, 140
  sign test, 62
  Wilcoxon signed-ranks, 62

binary data
  one-sample test, 49–55
binomial test, 49–50
  example: pilot study for new drug, 50
bivariate data
  measures of association, 166–167
blocked comparisons, 95
BY (keyword)
  NPAR TESTS command, 202

categorical data
  assumptions, 12
categorical variables, 135
CIN (keyword)
  CROSSTABS command, 199
  NPAR TESTS command, 200
class variables, 135
Cochran's Q test, 108–111
  example:cross-over clinical trial, 109–111
  when to use, 96
Cohen's kappa. *See* Kappa
confidence levels
  specifying, 7
contingency coefficients
  measures of association, 185, 185–188
contingency tables. *See* r x c contingency tables
continuous data
  assumptions, 12
continuous variables, 135
correlations
  Pearson's product-moment correlation coefficient, 172–174
  Spearman's rank-order correlation coefficient, 174–176
Cramer's *V*
  example, 187–188
  measures of association, 185–188
CROSSTABS (command), 199
  new syntax, 199
Crosstabs procedure, 199
  asymptotic *p* value, 7
  confidence levels, 7
  contingency coefficients, 185
  exact *p*value, 8
  exact statistics, 6–8
  Fisher's exact test, 141
  gamma, 171
  Goodman and Kruskal's tau, 185
  Kendall's tau-*b*, 171
  Kendall's tau-*c*, 171
  likelihood-ratio test, 141
  linear-by-linear association test, 155
  Monte Carlo *p* value, 7
  Pearson chi-square test, 141

Pearson's product moment correlation coefficient, 171
samples, 7
Somers' *d*, 171
Spearman's rank-order correlation coefficient, 171
time limit, 8
uncertainty coefficient, 185
crosstabulated data
measures of association, 165–167
crosstabulation, 199
See also Crosstabs procedure

data sets
small, 30
sparse, 36–37
tied, 31–34
unbalanced, 35
doubly ordered contingency tables, 135
doubly ordered contingency tables. *See also* r x c contingency tables

EXACT (keyword)
CROSSTABS command, 199
NPAR TESTS command, 200
exact method, 1–3
exact one-sided *p* value
K independent samples, 134
exact one-sided *p* value
Jonckheere-Terpstra test, 159
linear-by-linear association test, 162
Mann-Whitney test, 82
McNemar test, 69
runs test, 92
exact *p* value, 12, 16
defined, 1
example: fire figher data, 1–3
obtaining, 8
r x c tables, 136
when to use, 24
exact statistics
obtaining, 6–8
exact tests
memory limits, 8
setting time limit, 8
when to use, 4
exact two-sided *p* value
K independent samples, 134
median test, 124
exact two-sided *p* value
Jonckheere-Terpstra test, 160
K related samples, 99
Kolmogorov-Smirnov, 88
linear-by-linear assocation test, 162
Mann-Whitney test, 82
McNemar test, 69
measures of agreement, 168
nominal data, 168
ordinal data, 168
r x c tables, 138
runs test, 52

Fisher's exact test, 147–148
example: 2 x 2 table, 18–24
example: tea-tasting experiment, 18–24
when to use, 141
Friedman's test, 101–104
example: effect of hypnosis, 102–104
when to use, 96
full multinomial sampling, 137

gamma, 171
example: smoking habit data, 183–184
measures of association, 183–184
Goodman and Kruskal's tau
example: party preference data, 189–191
measures of association, 185, 188–191

independent samples, 75–94
Jonckheere-Terpstra test, 114, 131–134
when to use each test, 76

Jonchkeere-Terpstra test
example: space shuttle O-ring incidents, 132–134
Jonckheere-Terpstra test
asymptotic one-sided *p* value, 159
asymptotic two-sided *p* value, 159
exact one-sided *p* value, 159

exact two-sided *p* value, 160
example: dose-response data, 157–160
in Tests for Several Independent Samples procedure, 202
r x c contingency tables, 156–160
when to use, 115, 156
J-T (subcommand)
NPAR TESTS command, 202

K independent samples tests, 113–134
Jonckheere-Terpstra test, 131–134
Kruskal-Wallis test, 127–130
median test, 122–127
when to use, 114–115
K related samples tests, 95–111
Cochran's Q, 108–111
Friedman's, 101–104
Kendall's W, 104–107
when to use, 96
kappa
example:student teacher ratings, 193–195
measures of agreement, 193–195
Kendall's coefficient of concordance. See Kendall's *W*
Kendall's tau
example: smoking habit data, 180–182
measures of association, 177–182
Kendall's tau-*b*, 171
Kendall's tau-*c*, 171
Kendall's W test, 104–107
example: attendance at annual meeting, 105–107
example: relationship to Spearman's R, 107
when to use, 96
Kolmogorov-Smirnov test, 87–91
example: effectiveness of vitamin C, 90–91
example:diastolic blood pressure data, 31–34
when to use, 76
Kruskal-Wallis test, 149–153
example: hematologic toxicity data, 129–130
example: tumor regression rates, 150–153
when to use, 115, 143, 149

likelihood ratio test
example:sports activity data, 25–27
likelihood-ratio test, 145–147
when to use, 141
linear-by-linear association test
exact one-sided *p* value, 162
exact two-sided *p* value, 162
example: dose-response data, 161
example:alcohol and birth defect data, 35
r x c contingency tables, 161–164
when to use, 156
location-shift alternatives, 115

Mann-Whitney test, 80–86
example: blood pressure data, 84–86
when to use, 76
Mantel-Haenszel test. *See* linear-by-linear association test
marginal homogeneity test, 71–73
example: matched-case control study, 71–72
example: Pap-smear classification, 72–73
in Two-Related-Samples Tests procedure, 201–202
when to use, 58
MC (keyword)
CROSSTABS command, 199
NPAR TESTS command, 200
McNemar test, 68–70
exact one-sided *p* value, 69
exact two-sided *p* value, 69
example: voters' preference, 70
when to use, 58
measures of agreement
exact two-sided *p* value, 168
kappa, 193–195
measures of association
asymptotic *p* values, 169
bivariate data, 166–167
contingency coefficients, 185, 185–188
Cramer's *V*, 185–188
crosstabulated data, 165–167
exact *p* values, 168–169
gamma, 183–184
Goodman and Kruskal's tau, 188–191
introduction, 165–170
Kendall's tau, 177–182
Kendall's *W*, 171
Monte Carlo *p* values, 169
nominal data, 185–191
ordinal data, 171–184

p values, 168–170
Pearson's product-moment correlation coefficient, 171, 172–174
phi, 185–188
point estimates, 168
proportional reduction in prediction error, 188–191
proportional reduction in predictive error, 185
Somers' *d*, 177–182
Spearman's rank-order correlation coefficient, 171, 174–176
uncertainty coefficient, 189–191
median test, 122–127
example: hematologic toxicity data, 125–127
when to use, 115
memory limits
exact tests, 8
METHOD (subcommand)
CROSSTABS command, 199
NPAR TESTS command, 200–201, 202
MH (subcommand)
NPAR TESTS command, 201–202
Monte Carlo method, 3
defined, 3
example:fire figher data, 3
random number seed, 8–9
Monte Carlo one-sided *p* value
sign test, 63
Wilcoxon signed-ranks test, 63
Monte Carlo *p* value
obtaining, 7
when to use, 24–29
Monte Carlo *p* values
measures of association, 169
Monte Carlo two-sided *p* value
K independent samples, 120
median test, 124
Monte Carlo two-sided *p* value
K related samples, 100
Kolmogorov-Smirnov, 88
Mann-Whitney test, 83
r x c tables, 139
sign test, 64
Wilcoxon signed-ranks test, 64

nominal data
contingency coefficients, 185–188
Cramer's *V*, 185–188
exact two-sided *p* values, 168
Goodman and Kruskal's tau, 188–191
phi, 185–188
proportional reduction in prediction error, 188–191
uncertainty coefficient, 189–191
nominal variables, 135
nonparametric tests
assumptions, 12
asymptotic *p* value, 7
binomial, 49
Cochran's Q, 95
confidence levels, 7
exact *p*value, 8
exact statistics, 6–8
Friedman's, 95
Jonckheere-Terpstra test, 114, 155
Kendall's W, 95
Kolmogorov-Smirnov, 75
Kruskal-Wallis, 114, 149
Mann-Whitney test, 75
marginal homogeneity, 57
McNemar, 57
median test, 114
Monte Carlo *p* value, 7
new syntax, 200
new tests, 8
runs, 49, 75
samples, 7
sign, 57
time limit, 8
two-related samples, 57
Wald-Wolfowitz runs test, 75
Wilcoxon signed-ranks, 57
NPAR TESTS (command), 200–202
J-T subcommand, 202
METHOD subcommand, 200–201
MH subcommand, 201–202
new syntax, 200
pairing variables, 201

observed r x c tables, 135–136
computing exact *p* value for, 136
one-sample tests
binary data, 49–55
runs test, 51–55
one-sided *p* value
K independent samples, 120, 122

one-sided *p* value
binomial test, 50
Mann-Whitney test, 82, 84
McNemar test, 69
runs test, 92
sign test, 62, 63
Wilcoxon signed-ranks test, 62, 63
ordered alternatives, 115
ordered variables, 135
ordinal data
exact two-sided *p* values, 168
gamma, 183–184
Kendall's tau, 177–182
measures of association, 171–184
Pearson's product-moment correlation coefficient, 172–174
Somers' *d*, 177–182
Spearman's rank-order correlation coefficient, 174–176

p value
choosing a method, 22–37
hypothesis testing, 11–14
in two-sample tests, 80
measures of association, 168–170
p value. *See also* one-sided *p* value
p value. *See also* two-sided *p* value.
PAIRED (keyword)
NPAR TESTS command, 201
paired samples, 57–73
when to use each test, 58
Pearson chi-square
example: 3 x 4 table, 14–18
example: fire figher data, 14–18
example: sparse contingency table, 12–14
example: sports activity data, 36–37
Pearson chi-square test, 138, 144–145
when to use, 141
Pearson's product-moment correlation coefficient
example:social striving data, 30, 172–174
measures of association, 172–174
phi
example, 187–188
measures of association, 185–188
point estimates
measures of assocation, 168
Poisson sampling, 137
product multinomial sampling, 137, 143
proportional reduction in prediction error
measures of association, 185, 188–191
proportional reduction in prediction error. *See also* Goodman and Kruskal tau
proportional reduction in prediction error. *See also* uncertainty coefficient

r x c contingency tables
doubly ordered, 155–164
example: oral lesions data, 143–144
Jonckheere-Tepstra test, 156–160
Kruskal-Wallis test, 149–153
linear-by-linear association test, 161–164
observed, 135–136
reference sets for, 136
singly ordered, 149–153
tests on, 135–140
unordered, 141–148
random number seed, 8–9
reference sets, 16–17, 21, 137
for r x c tables, 136
runs test, 51–55, 91–94
example: children's aggression scores, 53–54
example: discrimination against female workers, 92–94
example: small data set, 54–55
when to use, 76

samples
Monte Carlo method, 7
SAMPLES (keyword)
NPAR TESTS command, 200
sampling
full multinomial, 137
Poisson, 137
product multinomial, 137
sign test, 59–67
when to use, 58
singly ordered contingency tables, 135
singly ordered contingency tables. *See also* r x c contingency tables
Somers' *d*, 171, 177–182

example: smoking habit data, 180–182
measures of association, 177–182
Spearman's rank-order correlation coefficient
example: social striving data, 175–176
measures of association, 174–176

test statistics
defining for r x c tables, 138
Tests for Several Independent Samples procedure, 200–202
grouping variables, 202
time limit
setting for exact tests, 8
TIMER (keyword)
NPAR TESTS command, 200
Two-Related-Samples Tests procedure, 201–202
two-sample tests
independent samples, 75–94
Kolmogorov-Smirnov, 87–91
Mann-Whitney, 80–86
marginal homogeneity, 71–73
McNemar, 68–70
median, 94
paired samples, 57–73
runs, 91–94
sign, 59–67
Wilcoxon signed-ranks, 59–67
two-sided *p* value
K independent samples, 115, 120, 121
median test, 124
two-sided *p* value
binomial test, 50
K related samples, 99, 101
Kolmogorov-Smirnov, 88
Mann-Whitney test, 82, 84
McNemar test, 69
r x c tables, 138, 140
runs test, 52
sign test, 62, 64
Wilcoxon signed-ranks test, 62, 64

uncertainty coefficient
example: party preference data, 189–191
measures of association, 185, 189–191
unordered continous contingency tables, 135
unordered r x c contingency tables
See also r x c contingency tables

Wald-Wolfowitz. *See* runs test
Wilcoxon rank-sum test, 11
Wilcoxon signed-rank test, 11
Wilcoxon signed-ranks test, 59–67
example: AZT for AIDS, 64–67
mid-ranks, 60
permutational distribution, 60
when to use, 58
Wilcoxon-Mann-Whitney test. *See* Mann-Whitney test
WITH (keyword)
NPAR TESTS command, 201