Nonparametric Statistical Inference

Fourth Edition, Revised and Expanded

Jean Dickinson Gibbons
Subhabrata Chakraborti
The University of Alabama
Tuscaloosa, Alabama, U.S.A.

MARCEL DEKKER, INC. NEW YORK • BASEL

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

ISBN: 0-8247-4052-1

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-260-6300; fax: 41-61-260-6333

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to Special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

STATISTICS: Textbooks and Monographs

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

1. The Generalized Jackknife Statistic, *H. L. Gray and W. R. Schucany*
2. Multivariate Analysis, *Anant M. Kshirsagar*
3. Statistics and Society, *Walter T. Federer*
4. Multivariate Analysis: A Selected and Abstracted Bibliography, 1957–1972, *Kocherlakota Subrahmaniam and Kathleen Subrahmaniam*
5. Design of Experiments: A Realistic Approach, *Virgil L. Anderson and Robert A. McLean*
6. Statistical and Mathematical Aspects of Pollution Problems, *John W. Pratt*
7. Introduction to Probability and Statistics (in two parts), Part I: Probability; Part II: Statistics, *Narayan C. Giri*
8. Statistical Theory of the Analysis of Experimental Designs, *J. Ogawa*
9. Statistical Techniques in Simulation (in two parts), *Jack P. C. Kleijnen*
10. Data Quality Control and Editing, *Joseph I. Naus*
11. Cost of Living Index Numbers: Practice, Precision, and Theory, *Kali S. Banerjee*
12. Weighing Designs: For Chemistry, Medicine, Economics, Operations Research, Statistics, *Kali S. Banerjee*
13. The Search for Oil: Some Statistical Methods and Techniques, *edited by D. B. Owen*
14. Sample Size Choice: Charts for Experiments with Linear Models, *Robert E. Odeh and Martin Fox*
15. Statistical Methods for Engineers and Scientists, *Robert M. Bethea, Benjamin S. Duran, and Thomas L. Boullion*
16. Statistical Quality Control Methods, *Irving W. Burr*
17. On the History of Statistics and Probability, *edited by D. B. Owen*
18. Econometrics, *Peter Schmidt*
19. Sufficient Statistics: Selected Contributions, *Vasant S. Huzurbazar (edited by Anant M. Kshirsagar)*
20. Handbook of Statistical Distributions, *Jagdish K. Patel, C. H. Kapadia, and D. B. Owen*
21. Case Studies in Sample Design, *A. C. Rosander*
22. Pocket Book of Statistical Tables, *compiled by R. E. Odeh, D. B. Owen, Z. W. Birnbaum, and L. Fisher*
23. The Information in Contingency Tables, *D. V. Gokhale and Solomon Kullback*
24. Statistical Analysis of Reliability and Life-Testing Models: Theory and Methods, *Lee J. Bain*
25. Elementary Statistical Quality Control, *Irving W. Burr*
26. An Introduction to Probability and Statistics Using BASIC, *Richard A. Groeneveld*
27. Basic Applied Statistics, *B. L. Raktoe and J. J. Hubert*
28. A Primer in Probability, *Kathleen Subrahmaniam*
29. Random Processes: A First Look, *R. Syski*
30. Regression Methods: A Tool for Data Analysis, *Rudolf J. Freund and Paul D. Minton*
31. Randomization Tests, *Eugene S. Edgington*
32. Tables for Normal Tolerance Limits, Sampling Plans and Screening, *Robert E. Odeh and D. B. Owen*
33. Statistical Computing, *William J. Kennedy, Jr., and James E. Gentle*
34. Regression Analysis and Its Application: A Data-Oriented Approach, *Richard F. Gunst and Robert L. Mason*
35. Scientific Strategies to Save Your Life, *I. D. J. Bross*
36. Statistics in the Pharmaceutical Industry, *edited by C. Ralph Buncher and Jia-Yeong Tsay*
37. Sampling from a Finite Population, *J. Hajek*
38. Statistical Modeling Techniques, *S. S. Shapiro and A. J. Gross*
39. Statistical Theory and Inference in Research, *T. A. Bancroft and C.-P. Han*
40. Handbook of the Normal Distribution, *Jagdish K. Patel and Campbell B. Read*
41. Recent Advances in Regression Methods, *Hrishikesh D. Vinod and Aman Ullah*
42. Acceptance Sampling in Quality Control, *Edward G. Schilling*
43. The Randomized Clinical Trial and Therapeutic Decisions, *edited by Niels Tygstrup, John M. Lachin, and Erik Juhl*

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

44. Regression Analysis of Survival Data in Cancer Chemotherapy, *Walter H Carter, Jr, Galen L Wampler, and Donald M Stablein*
45. A Course in Linear Models, *Anant M Kshirsagar*
46. Clinical Trials: Issues and Approaches, *edited by Stanley H Shapiro and Thomas H Louis*
47. Statistical Analysis of DNA Sequence Data, *edited by B S Weir*
48. Nonlinear Regression Modeling: A Unified Practical Approach, *David A Ratkowsky*
49. Attribute Sampling Plans, Tables of Tests and Confidence Limits for Proportions, *Robert E Odeh and D B Owen*
50. Experimental Design, Statistical Models, and Genetic Statistics, *edited by Klaus Hinkelmann*
51. Statistical Methods for Cancer Studies, *edited by Richard G Cornell*
52. Practical Statistical Sampling for Auditors, *Arthur J. Wilburn*
53 Statistical Methods for Cancer Studies, *edited by Edward J Wegman and James G Smith*
54 Self-Organizing Methods in Modeling GMDH Type Algorithms, *edited by Stanley J. Farlow*
55 Applied Factorial and Fractional Designs, *Robert A McLean and Virgil L. Anderson*
56 Design of Experiments Ranking and Selection, *edited by Thomas J Santner and Ajit C Tamhane*
57 Statistical Methods for Engineers and Scientists: Second Edition, Revised and Expanded, *Robert M. Bethea, Benjamin S Duran, and Thomas L Boullion*
58 Ensemble Modeling. Inference from Small-Scale Properties to Large-Scale Systems, *Alan E Gelfand and Crayton C Walker*
59 Computer Modeling for Business and Industry, *Bruce L Bowerman and Richard T O'Connell*
60 Bayesian Analysis of Linear Models, *Lyle D Broemeling*
61. Methodological Issues for Health Care Surveys, *Brenda Cox and Steven Cohen*
62 Applied Regression Analysis and Experimental Design, *Richard J Brook and Gregory C Arnold*
63 Statpal: A Statistical Package for Microcomputers—PC-DOS Version for the IBM PC and Compatibles, *Bruce J Chalmer and David G Whitmore*
64. Statpal: A Statistical Package for Microcomputers—Apple Version for the II, II+, and IIe, *David G Whitmore and Bruce J. Chalmer*
65. Nonparametric Statistical Inference. Second Edition, Revised and Expanded, *Jean Dickinson Gibbons*
66 Design and Analysis of Experiments, *Roger G Petersen*
67. Statistical Methods for Pharmaceutical Research Planning, *Sten W Bergman and John C Gittins*
68. Goodness-of-Fit Techniques, *edited by Ralph B D'Agostino and Michael A. Stephens*
69. Statistical Methods in Discrimination Litigation, *edited by D H Kaye and Mikel Aickin*
70. Truncated and Censored Samples from Normal Populations, *Helmut Schneider*
71. Robust Inference, *M L Tiku, W Y Tan, and N Balakrishnan*
72. Statistical Image Processing and Graphics, *edited by Edward J. Wegman and Douglas J DePriest*
73. Assignment Methods in Combinatorial Data Analysis, *Lawrence J Hubert*
74. Econometrics and Structural Change, *Lyle D Broemeling and Hiroki Tsurumi*
75. Multivariate Interpretation of Clinical Laboratory Data, *Adelin Albert and Eugene K Harris*
76. Statistical Tools for Simulation Practitioners, *Jack P C Kleijnen*
77. Randomization Tests: Second Edition, *Eugene S Edgington*
78 A Folio of Distributions A Collection of Theoretical Quantile-Quantile Plots, *Edward B Fowlkes*
79. Applied Categorical Data Analysis, *Daniel H Freeman, Jr*
80. Seemingly Unrelated Regression Equations Models: Estimation and Inference, *Virendra K Srivastava and David E A Giles*

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

81. Response Surfaces: Designs and Analyses, *Andre I. Khuri and John A. Cornell*
82. Nonlinear Parameter Estimation: An Integrated System in BASIC, *John C. Nash and Mary Walker-Smith*
83. Cancer Modeling, *edited by James R. Thompson and Barry W. Brown*
84. Mixture Models: Inference and Applications to Clustering, *Geoffrey J. McLachlan and Kaye E. Basford*
85. Randomized Response. Theory and Techniques, *Arijit Chaudhuri and Rahul Mukerjee*
86 Biopharmaceutical Statistics for Drug Development, *edited by Karl E. Peace*
87. Parts per Million Values for Estimating Quality Levels, *Robert E Odeh and D B. Owen*
88. Lognormal Distributions: Theory and Applications, *edited by Edwin L. Crow and Kunio Shimizu*
89. Properties of Estimators for the Gamma Distribution, *K O. Bowman and L. R. Shenton*
90. Spline Smoothing and Nonparametric Regression, *Randall L Eubank*
91. Linear Least Squares Computations, *R W Farebrother*
92. Exploring Statistics, *Damaraju Raghavarao*
93. Applied Time Series Analysis for Business and Economic Forecasting, *Sufi M Nazem*
94 Bayesian Analysis of Time Series and Dynamic Models, *edited by James C. Spall*
95. The Inverse Gaussian Distribution: Theory, Methodology, and Applications, *Raj S Chhikara and J. Leroy Folks*
96. Parameter Estimation in Reliability and Life Span Models, *A. Clifford Cohen and Betty Jones Whitten*
97. Pooled Cross-Sectional and Time Series Data Analysis, *Terry E Dielman*
98. Random Processes: A First Look, Second Edition, Revised and Expanded, *R. Syski*
99. Generalized Poisson Distributions: Properties and Applications, *P C. Consul*
100. Nonlinear L_p-Norm Estimation, *Rene Gonin and Arthur H Money*
101. Model Discrimination for Nonlinear Regression Models, *Dale S. Borowiak*
102. Applied Regression Analysis in Econometrics, *Howard E. Doran*
103. Continued Fractions in Statistical Applications, *K. O. Bowman and L. R. Shenton*
104 Statistical Methodology in the Pharmaceutical Sciences, *Donald A. Berry*
105. Experimental Design in Biotechnology, *Perry D. Haaland*
106. Statistical Issues in Drug Research and Development, *edited by Karl E Peace*
107. Handbook of Nonlinear Regression Models, *David A. Ratkowsky*
108. Robust Regression: Analysis and Applications, *edited by Kenneth D Lawrence and Jeffrey L Arthur*
109. Statistical Design and Analysis of Industrial Experiments, *edited by Subir Ghosh*
110. *U*-Statistics: Theory and Practice, *A J Lee*
111. A Primer in Probability: Second Edition, Revised and Expanded, *Kathleen Subrahmaniam*
112. Data Quality Control: Theory and Pragmatics, *edited by Gunar E Liepins and V. R R. Uppuluri*
113. Engineering Quality by Design: Interpreting the Taguchi Approach, *Thomas B Barker*
114 Survivorship Analysis for Clinical Studies, *Eugene K. Harris and Adelin Albert*
115. Statistical Analysis of Reliability and Life-Testing Models: Second Edition, *Lee J. Bain and Max Engelhardt*
116. Stochastic Models of Carcinogenesis, *Wai-Yuan Tan*
117. Statistics and Society: Data Collection and Interpretation, Second Edition, Revised and Expanded, *Walter T. Federer*
118. Handbook of Sequential Analysis, *B K. Ghosh and P. K. Sen*
119 Truncated and Censored Samples: Theory and Applications, *A. Clifford Cohen*
120. Survey Sampling Principles, *E. K. Foreman*
121. Applied Engineering Statistics, *Robert M. Bethea and R. Russell Rhinehart*
122. Sample Size Choice: Charts for Experiments with Linear Models: Second Edition, *Robert E. Odeh and Martin Fox*
123. Handbook of the Logistic Distribution, *edited by N Balakrishnan*
124. Fundamentals of Biostatistical Inference, *Chap T. Le*
125. Correspondence Analysis Handbook, *J.-P Benzécri*

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

126. Quadratic Forms in Random Variables: Theory and Applications, *A. M. Mathai and Serge B. Provost*
127. Confidence Intervals on Variance Components, *Richard K. Burdick and Franklin A. Graybill*
128. Biopharmaceutical Sequential Statistical Applications, *edited by Karl E. Peace*
129. Item Response Theory: Parameter Estimation Techniques, *Frank B. Baker*
130. Survey Sampling: Theory and Methods, *Arijit Chaudhuri and Horst Stenger*
131. Nonparametric Statistical Inference: Third Edition, Revised and Expanded, *Jean Dickinson Gibbons and Subhabrata Chakraborti*
132. Bivariate Discrete Distribution, *Subrahmaniam Kocherlakota and Kathleen Kocherlakota*
133. Design and Analysis of Bioavailability and Bioequivalence Studies, *Shein-Chung Chow and Jen-pei Liu*
134. Multiple Comparisons, Selection, and Applications in Biometry, *edited by Fred M. Hoppe*
135. Cross-Over Experiments: Design, Analysis, and Application, *David A. Ratkowsky, Marc A. Evans, and J. Richard Alldredge*
136. Introduction to Probability and Statistics: Second Edition, Revised and Expanded, *Narayan C. Giri*
137. Applied Analysis of Variance in Behavioral Science, *edited by Lynne K. Edwards*
138. Drug Safety Assessment in Clinical Trials, *edited by Gene S. Gilbert*
139. Design of Experiments: A No-Name Approach, *Thomas J. Lorenzen and Virgil L. Anderson*
140. Statistics in the Pharmaceutical Industry: Second Edition, Revised and Expanded, *edited by C. Ralph Buncher and Jia-Yeong Tsay*
141. Advanced Linear Models: Theory and Applications, *Song-Gui Wang and Shein-Chung Chow*
142. Multistage Selection and Ranking Procedures: Second-Order Asymptotics, *Nitis Mukhopadhyay and Tumulesh K. S. Solanky*
143. Statistical Design and Analysis in Pharmaceutical Science: Validation, Process Controls, and Stability, *Shein-Chung Chow and Jen-pei Liu*
144. Statistical Methods for Engineers and Scientists: Third Edition, Revised and Expanded, *Robert M. Bethea, Benjamin S. Duran, and Thomas L. Boullion*
145. Growth Curves, *Anant M. Kshirsagar and William Boyce Smith*
146. Statistical Bases of Reference Values in Laboratory Medicine, *Eugene K. Harris and James C. Boyd*
147. Randomization Tests: Third Edition, Revised and Expanded, *Eugene S. Edgington*
148. Practical Sampling Techniques: Second Edition, Revised and Expanded, *Ranjan K. Som*
149. Multivariate Statistical Analysis, *Narayan C. Giri*
150. Handbook of the Normal Distribution: Second Edition, Revised and Expanded, *Jagdish K. Patel and Campbell B. Read*
151. Bayesian Biostatistics, *edited by Donald A. Berry and Dalene K. Stangl*
152. Response Surfaces: Designs and Analyses, Second Edition, Revised and Expanded, *André I. Khuri and John A. Cornell*
153. Statistics of Quality, *edited by Subir Ghosh, William R. Schucany, and William B. Smith*
154. Linear and Nonlinear Models for the Analysis of Repeated Measurements, *Edward F. Vonesh and Vernon M. Chinchilli*
155. Handbook of Applied Economic Statistics, *Aman Ullah and David E. A. Giles*
156. Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators, *Marvin H. J. Gruber*
157. Nonparametric Regression and Spline Smoothing: Second Edition, *Randall L. Eubank*
158. Asymptotics, Nonparametrics, and Time Series, *edited by Subir Ghosh*
159. Multivariate Analysis, Design of Experiments, and Survey Sampling, *edited by Subir Ghosh*

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Additional Volumes in Preparation

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

To the memory of my parents,
John and Alice,
And to my husband, John S. Fielden
J.D.G.

To my parents,
Himangshu and Pratima,
And to my wife Anuradha, and son, Siddhartha Neil
S.C.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Preface to the Fourth Edition

This book was first published in 1971 and last revised in 1992. During the span of over 30 years, it seems fair to say that the book has made a meaningful contribution to the teaching and learning of nonparametric statistics. We have been gratified by the interest and the comments from our readers, reviewers, and users. These comments and our own experiences have resulted in many corrections, improvements, and additions.

We have two main goals in this revision: We want to bring the material covered in this book into the 21st century, and we want to make the material more user friendly.

With respect to the first goal, we have added new materials concerning the quantiles, the calculation of exact power and simulated power, sample size determination, other goodness-of-fit tests, and multiple comparisons. These additions will be discussed in more detail later. We have added and modified examples and included exact solutions done by hand and modern computer solutions using MINITAB,* SAS, STATXACT, and SPSS. We have removed most of the computer solutions to previous examples using BMDP, SPSSX, Execustat, or IMSL, because they seem redundant and take up too much valuable space. We have added a number of new references but have made no attempt to make the references comprehensive on some current minor refinements of the procedures covered. Given the sheer volume of the literature, preparing a comprehensive list of references on the subject of nonparametric statistics would truly be a challenging task. We apologize to the authors whose contributions could not be included in this edition.

With respect to our second goal, we have completely revised a number of sections and reorganized some of the materials, more fully integrated the applications with the theory, given tabular guides for applications of tests and confidence intervals, both exact and approximate, placed more emphasis on reporting results using P values, added some new problems, added many new figures and titled all figures and tables, supplied answers to almost all the problems, increased the number of numerical examples with solutions, and written concise but detailed summaries for each chapter. We think the problem answers should be a major plus, something many readers have requested over the years. We have also tried to correct errors and inaccuracies from previous editions.

In Chapter 1, we have added Chebyshev's inequality, the Central Limit Theorem, and computer simulations, and expanded the listing of probability functions, including the multinomial distribution and the relation between the beta and gamma functions. Chapter 2 has been completely reorganized, starting with the quantile function and the empirical distribution function (edf), in an attempt to motivate the reader to see the importance of order statistics. The relation between rank and the edf is explained. The tests and confidence intervals for quantiles have been moved to Chapter 5 so that they are discussed along with other one-sample and paired-sample procedures, namely, the sign test and signed rank test for the median. New discussions of exact power, simulated power, and sample size determination, and the discussion of rank tests in Chapter 5 of the previous edition are also included here. Chapter 4, on goodness-of-fit tests, has been expanded to include Lilliefors's test for the exponential distribution,

* MINITAB is a trademark of Minitab Inc. in the United States and other countries and is used herein with permission of the owner (on the Web at www.minitab.com).

computation of normal probability plots, and visual analysis of goodness of fit using *P*-*P* and *Q*-*Q* plots.

The new Chapter 6, on the general two-sample problem, defines "stochastically larger" and gives numerical examples with exact and computer solutions for all tests. We include sample size determination for the Mann-Whitney-Wilcoxon test. Chapters 7 and 8 are the previous-edition Chapters 8 and 9 on linear rank tests for the location and scale problems, respectively, with numerical examples for all procedures. The method of positive variables to obtain a confidence interval estimate of the ratio of scale parameters when nothing is known about location has been added to Chapter 8, along with a much needed summary.

Chapters 10 and 12, on tests for k samples, now include multiple comparisons procedures. The materials on nonparametric correlation in Chapter 11 have been expanded to include the interpretation of Kendall's tau as a coefficient of disarray, the Student's t approximation to the distribution of Spearman's rank correlation coefficient, and the definitions of Kendall's tau a, tau b and the Goodman-Kruskal coefficient. Chapter 14, a new chapter, discusses nonparametric methods for analyzing count data. We cover analysis of contingency tables, tests for equality of proportions, Fisher's exact test, McNemar's test, and an adaptation of Wilcoxon's rank-sum test for tables with ordered categories.

Bergmann, Ludbrook, and Spooren (2000) warn of possible meaningful differences in the outcomes of P values from different statistical packages. These differences can be due to the use of exact versus asymptotic distributions, use or nonuse of a continuity correction, or use or nonuse of a correction for ties. The output seldom gives such details of calculations, and even the "Help" facility and the manuals do not always give a clear description or documentation of the methods used to carry out the computations. Because this warning is quite valid, we tried to explain to the best of our ability any differences between our hand calculations and the package results for each of our examples.

As we said at the beginning, it has been most gratifying to receive very positive remarks, comments, and helpful suggestions on earlier editions of this book and we sincerely thank many readers and colleagues who have taken the time. We would like to thank Minitab, Cytel, and Statsoft for providing complimentary copies of their software. The popularity of nonparametric statistics must depend, to some extent, on the availability of inexpensive and user-friendly software. Portions of MINITAB Statistical Software input and output in this book are reprinted with permission of Minitab Inc.

Many people have helped, directly and indirectly, to bring a project of this magnitude to a successful conclusion. We are thankful to the University of Alabama and to the Department of Information Systems, Statistics and Management Science for providing an environment conducive to creative work and for making some resources available. In particular, Heather Davis has provided valuable assistance with typing. We are indebted to Clifton D. Sutton of George Mason University for pointing out errors in the first printing of the third edition. These have all been corrected. We are grateful to Joseph Stubenrauch, Production Editor at Marcel Dekker for giving us excellent editorial assistance. We also thank the reviewers of the third edition for their helpful comments and suggestions. These include Jones (1993), Prvan (1993), and Ziegel (1993). Ziegel's review in *Technometrics* stated, "This is *the* book for all statisticians and students in statistics who need to learn nonparametric statistics— . . .. I am grateful that the author decided that one more edition could already improve a fine package." We sincerely hope that Mr. Ziegel and others will agree that this fine package has been improved in scope, readability, and usability.

Jean Dickinson Gibbons
Subhabrata Chakraborti

Preface to the Third Edition

The third edition of this book includes a large amount of additions and changes. The additions provide a broader coverage of the nonparametric theory and methods, along with the tables required to apply them in practice. The primary change in presentation is an integration of the discussion of theory, applications, and numerical examples of applications. Thus the book has been returned to its original fourteen chapters with illustrations of practical applications following the theory at the appropriate place within each chapter. In addition, many of the hand-calculated solutions to these examples are verified and illustrated further by showing the solutions found by using one or more of the frequently used computer packages. When the package solutions are not equivalent, which happens frequently because most of the packages use approximate sampling distributions, the reasons are discussed briefly. Two new packages have recently been developed exclusively for nonparametric methods—NONPAR: Nonparametric Statistics Package and STATXACT: A Statistical Package for Exact

Nonparametric Inference. The latter package claims to compute exact P values. We have not used them but still regard them as a welcome addition.

Additional new material is found in the problem sets at the end of each chapter. Some of the new theoretical problems request verification of results published in journals about inference procedures not covered specifically in the text. Other new problems refer to the new material included in this edition. Further, many new applied problems have been added.

The new topics that are covered extensively are as follows. In Chapter 2 we give more convenient expressions for the moments of order statistics in terms of the quantile function, introduce the empirical distribution function, and discuss both one-sample and two-sample coverages so that problems can be given relating to exceedance and precedence statistics. The rank von Neumann test for randomness is included in Chapter 3 along with applications of runs tests in analyses of time series data. In Chapter 4 on goodness-of-fit tests, Lilliefors's test for a normal distribution with unspecified mean and variance has been added.

Chapter 7 now includes discussion of the control median test as another procedure appropriate for the general two-sample problem. The extension of the control median test to k mutually independent samples is given in Chapter 11. Other new materials in Chapter 11 are nonparametric tests for ordered alternatives appropriate for data based on $k \geqslant 3$ mutually independent random samples. The tests proposed by Jonckheere and Terpstra are covered in detail. The problems relating to comparisons of treatments with a control or an unknown standard are also included here.

Chapter 13, on measures of association in multiple classifications, has an additional section on the Page test for ordered alternatives in k-related samples, illustration of the calculation of Kendall's tau for count data in ordered contingency tables, and calculation of Kendall's coefficient of partial correlation. Chapter 14 now includes calculations of asymptotic relative efficiency of more tests and also against more parent distributions.

For most tests covered, the corrections for ties are derived and discussions of relative performance are expanded. New tables included in the Appendix are the distributions of the Lilliefors's test for normality, Kendall's partial tau, Page's test for ordered alternatives in the two-way layout, the Jonckheere-Terpstra test for ordered alternatives in the one-way layout, and the rank von Neumann test for randomness.

This edition also includes a large number of additional references. However, the list of references is not by any means purported to be complete because the literature on nonparametric inference procedures is vast. Therefore, we apologize to those authors whose contributions were not included in our list of references.

As always in a new edition, we have attempted to correct previous errors and inaccuracies and restate more clearly the text and problems retained from previous editions. We have also tried to take into account the valuable suggestions for improvement made by users of previous editions and reviewers of the second edition, namely, Moore (1986), Randles (1986), Sukhatme (1987), and Ziegel (1988).

As with any project of this magnitude, we are indebted to many persons for help. In particular, we would like to thank Pat Coons and Connie Harrison for typing and Nancy Kao for help in the bibliography search and computer solutions to examples. Finally, we are indebted to the University of Alabama, particularly the College of Commerce and Business Administration, for partial support during the writing of this version.

Jean Dickinson Gibbons
Subhabrata Chakraborti

Preface to the Second Edition

A large number of books on nonparametric statistics have appeared since this book was published in 1971. The majority of them are oriented toward applications of nonparametric methods and do not attempt to explain the theory behind the techniques; they are essentially user's manuals, called cookbooks by some. Such books serve a useful purpose in the literature because non-parametric methods have such a broad scope of application and have achieved widespread recognition as a valuable technique for analyzing data, particularly data which consist of ranks or relative preferences and/or are small samples from unknown distributions. These books are generally used by nonstatisticians, that is, persons in subject-matter fields. The more recent books that are oriented toward theory are Lehmann (1975), Randles and Wolfe (1979), and Pratt and Gibbons (1981).

A statistician needs to know about both the theory and methods of nonparametric statistical inference. However, most graduate programs in statistics can afford to offer either a theory course or a methods course, but not both. The first edition of this book was frequently used for the theory course; consequently, the students were forced to learn applications on their own time.

This second edition not only presents the theory with corrections from the first edition, it also offers substantial practice in problem solving. Chapter 15 of this edition includes examples of applications of those techniques for which the theory has been presented in Chapters 1 to 14. Many applied problems are given in this new chapter; these problems involve real research situations from all areas of social, behavioral, and life sciences, business, engineering, and so on. The Appendix of Tables at the end of this new edition gives those tables of exact sampling distributions that are necessary for the reader to understand the examples given and to be able to work out the applied problems. To make it easy for the instructor to cover applications as soon as the relevant theory has been presented, the sections of Chapter 15 follow the order of presentation of theory. For example, after Chapter 3 on tests based on runs is completed, the next assignment can be Section 15.3 on applications of tests based on runs and the accompanying problems at the end of that section. At the end of the Chapter 15 there are a large number of review problems arranged in random order as to type of applications so that the reader can obtain practice in selecting the appropriate nonparametric technique to use in a given situation.

While the first edition of this book received considerable acclaim, several reviewers felt that applied numerical examples and expanded problem sets would greatly enhance its usefulness as a textbook. This second edition incorporates these and other recommendations. The author wishes to acknowledge her indebtedness to the following reviewers for helping to make this revised and expanded edition more accurate and useful for students and researchers: Dudewicz and Geller (1972), Johnson (1973), Klotz (1972), and Noether (1972).

In addition to these persons, many users of the first edition have written or told me over the years about their likes and/or dislikes regarding the book and these have all been gratefully received and considered for incorporation in this edition. I would also like to express my gratitude to Donald B. Owen for suggesting and encouraging this kind of revision, and to the Board of Visitors of the University of Alabama for partial support of this project.

Jean Dickinson Gibbons

Preface to the First Edition

During the last few years many institutions offering graduate programs in statistics have experienced a demand for a course devoted exclusively to a survey of nonparametric techniques and their justifications. This demand has arisen both from their own majors and from majors in social science or other quantitatively oriented fields such as psychology, sociology, or economics. Although the basic statistics courses often include a brief description of some of the better-known and simpler nonparametric methods, usually the treatment is necessarily perfunctory and perhaps even misleading. Discussion of only a few techniques in a highly condensed fashion may leave the impression that nonparametric statistics consists of a "bundle of tricks" which are simply applied by following a list of instructions dreamed up by some genie as a panacea for all sorts of vague and ill-defined problems.

One of the deterrents to meeting this demand has been the lack of a suitable textbook in nonparametric techniques. Our experience at the University of Pennsylvania has indicated that an appropriate text would provide a theoretical but readable survey. Only a moderate amount of pure mathematical sophistication should be required so that the course would be comprehensible to a wide variety of graduate students and perhaps even some advanced undergraduates. The course should be available to anyone who has completed at least the rather traditional one-year sequence in probability and statistical inference at the level of Parzen, Mood and Graybill, Hogg and Craig, etc. The time allotment should be a full semester, or perhaps two semesters if outside reading in journal publications is desirable.

The texts presently available which are devoted exclusively to nonparametric statistics are few in number and seem to be predominantly either of the handbook style, with few or no justifications, or of the highly rigorous mathematical style. The present book is an attempt to bridge the gap between these extremes. It assumes the reader is well acquainted with statistical inference for the traditional parametric estimation and hypothesis-testing procedures, basic probability theory, and random-sampling distributions. The survey is not intended to be exhaustive, as the field is so extensive. The purpose of the book is to provide a compendium of some of the better-known nonparametric techniques for each problem situation. Those derivations, proofs, and mathematical details which are relatively easily grasped or which illustrate typical procedures in general nonparametric statistics are included. More advanced results are simply stated with references. For example, some of the asymptotic distribution theory for order statistics is derived since the methods are equally applicable to other statistical problems. However, the Glivenko Cantelli theorem is given without proof since the mathematics may be too advanced. Generally those proofs given are not mathematically rigorous, ignoring details such as existence of derivatives or regularity conditions. At the end of each chapter, some problems are included which are generally of a theoretical nature but on the same level as the related text material they supplement.

The organization of the material is primarily according to the type of statistical information collected and the type of questions to be answered by the inference procedures or according to the general type of mathematical derivation. For each statistic, the null distribution theory is derived, or when this would be too tedious, the procedure one could follow is outlined, or when this would be overly theoretical, the results are stated without proof. Generally the other relevant mathematical details necessary for nonparametric inference are also included. The purpose is to acquaint the reader with the mathematical

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

logic on which a test is based, those test properties which are essential for understanding the procedures, and the basic tools necessary for comprehending the extensive literature published in the statistics journals. The book is not intended to be a user's manual for the application of nonparametric techniques. As a result, almost no numerical examples or problems are provided to illustrate applications or elicit applied motivation. With the approach, reproduction of an extensive set of tables is not required.

The reader may already be acquainted with many of the nonparametric methods. If not, the foundations obtained from this book should enable anyone to turn to a user's handbook and quickly grasp the application. Once armed with the theoretical background, the user of nonparametric methods is much less likely to apply tests indiscriminately or view the field as a collection of simple prescriptions. The only insurance against misapplication is a thorough understanding. Although some of the strengths and weaknesses of the tests covered are alluded to, no definitive judgments are attempted regarding the relative merits of comparable tests. For each topic covered, some references are given which provide further information about the tests or are specifically related to the approach used in this book. These references are necessarily incomplete, as the literature is vast. The interested reader may consult Savage's "Bibliography" (1962).

I wish to acknowledge the helpful comments of the reviewers and the assistance provided unknowingly by the authors of other textbooks in the area of nonparametric statistics, particularly Gottfried E. Noether and James V. Bradley, for the approach to presentation of several topics, and Maurice G. Kendall, for much of the material on measures of association. The products of their endeavors greatly facilitated this project. It is a pleasure also to acknowledge my indebtedness to Herbert A. David, both as friend and mentor. His training and encouragement helped make this book a reality. Particular gratitude is also due to the Lecture Note Fund of the Wharton School, for typing assistance, and the Department of Statistics and Operations Research at the University of Pennsylvania for providing the opportunity and time to finish this manuscript. Finally, I thank my husband for his enduring patience during the entire writing stage.

Jean Dickinson Gibbons

Contents

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

1
Introduction and Fundamentals

1.1 INTRODUCTION

In many elementary statistics courses, the subject matter is somewhat arbitrarily divided into two categories, called descriptive and inductive statistics. *Descriptive statistics* usually relates only to the calculation or presentation of figures (visual or conceptual) to summarize or characterize a set of data. For such procedures, no assumptions are made or implied, and there is no question of legitimacy of techniques. The descriptive figures may be a mean, median, variance, range, histogram, etc. Each of these figures summarizes a set of numbers in its own unique way; each is a distinguishable and well-defined characterization of data. If such data constitute a random sample from a certain population, the sample represents the population in miniature and any set of descriptive statistics provides some information regarding this universe. The term *parameter* is generally employed to connote a characteristic of the population. A parameter is often an

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

unspecified constant appearing in a family of probability distributions, but the word can also be interpreted in a broader sense to include almost all descriptions of population characteristics within a family.

When sample descriptions are used to infer some information about the population, the subject is called *inductive statistics or statistical inference*. The two types of problems most frequently encountered here are estimation and tests of hypotheses. The factor which makes inference a scientific method, thereby differentiating it from mere guessing, is the ability to make evaluations or probability statements concerning the accuracy of an estimate or reliability of a decision. Unfortunately, such scientific evaluations cannot be made without some information regarding the probability distribution of the random variable relating to the sample description used in the inference procedure. This means that certain types of sample descriptions will be more popular than others, because of their distribution properties or mathematical tractability. The sample arithmetic mean is a popular figure for describing the characteristic of central tendency for many reasons but perhaps least of all because it is a mean. The unique position of the mean in inference stems largely from its "almost normal" distribution properties. If some other measure, say the sample median, had a property as useful as the central-limit theorem, surely it would share the spotlight as a favorite description of location.

The entire body of classical statistical-inference techniques is based on fairly specific assumptions regarding the nature of the underlying population distribution; usually its form and some parameter values must be stated. Given the right set of assumptions, certain test statistics can be developed using mathematics which is frequently elegant and beautiful. The derived distribution theory is qualified by certain prerequisite conditions, and therefore all conclusions reached using these techniques are exactly valid only so long as the assumptions themselves can be substantiated. In textbook problems, the requisite postulates are frequently just stated and the student practices applying the appropriate technique. However, in a real-world problem, everything does not come packaged with labels of population of origin. A decision must be made as to what population properties may judiciously be assumed for the model. If the reasonable assumptions are not such that the traditional techniques are applicable, the classical methods may be used and inference conclusions stated only with the appropriate qualifiers, e.g., "If the population is normal, then . . . ."

The mathematical statistician may claim that it is the users' problem to decide on the legitimacy of the postulates. Frequently in practice, those assumptions which are deemed reasonable by empirical evidence or past experience are not the desired ones, i.e., those for which a set of standard statistical techniques has been developed. Alternatively, the sample may be too small or previous experience too limited to determine what is a reasonable assumption. Or, if the researcher is a product of the "cookbook school" of statistics, his particular expertise being in the area of application, he may not understand or even be aware of the preconditions implicit in the derivation of the statistical technique. In any of these three situations, the result often is a substitution of blind faith for scientific method, either because of ignorance or with the rationalization that an approximately accurate inference based on recognized and accepted scientific techniques is better than no answer at all or a conclusion based on common sense or intuition.

An alternative set of techniques is available, and the mathematical bases for these procedures are the subject of this book. They may be classified as distribution-free and nonparametric procedures. In a *distribution-free* inference, whether for testing or estimation, the methods are based on functions of the sample observations whose corresponding random variable has a distribution which does not depend on the specific distribution function of the population from which the sample was drawn. Therefore, assumptions regarding the underlying population are not necessary. On the other hand, strictly speaking, the term *nonparametric test* implies a test for a hypothesis which is not a statement about parameter values. The type of statement permissible then depends on the definition accepted for the term parameter. If parameter is interpreted in the broader sense, the hypothesis can be concerned only with the form of the population, as in goodness-of-fit tests, or with some characteristic of the probability distribution of the sample data, as in tests of randomness and trend. Needless to say, distribution-free tests and nonparametric tests are not synonymous labels or even in the same spirit, since one relates to the distribution of the test statistic and the other to the type of hypothesis to be tested. A distribution-free test may be for a hypothesis concerning the median, which is certainly a population parameter within our broad definition of the term.

In spite of the inconsistency in nomenclature, we shall follow the customary practice and consider both types of tests as procedures in nonparametric inference, making no distinction between the two classifications. For the purpose of differentiation, the classical statistical techniques, whose justification in probability is based on specific assumptions about the population sampled, may be called *parametric methods*. This implies a definition of nonparametric statistics then as the treatment of either nonparametric types of inferences or analogies to standard statistical problems when specific distribution assumptions are replaced by very general assumptions and the analysis is based on some function of the sample observations whose sampling distribution can be determined without knowledge of the specific distribution function of the underlying population. The assumption most frequently required is simply that the population be continuous. More restrictive assumptions are sometimes made, e.g., that the population is symmetrical, but not to the extent that the distribution is specifically postulated. The information used in making nonparametric inferences generally relates to some function of the actual magnitudes of the random variables in the sample. For example, if the actual observations are replaced by their relative rankings within the sample and the probability distribution of some function of these sample ranks can be determined by postulating only very general assumptions about the basic population sampled, this function will provide a distribution-free technique for estimation or hypothesis testing. Inferences based on descriptions of these derived sample data may relate to whatever parameters are relevant and adaptable, such as the median for a location parameter. The nonparametric and parametric hypotheses are analogous, both relating to location, and identical in the case of a continuous and symmetrical population.

Tests of hypotheses which are not statements about parameter values have no counterpart in parametric statistics; and thus here nonparametric statistics provides techniques for solving new kinds of problems. On the other hand, a distribution-free test simply relates to a different approach to solving standard statistical problems, and therefore comparisons of the merits of the two types of techniques are relevant. Some of the more obvious general advantages of nonparametric-inference procedures can be appreciated even before our systematic study begins. Nonparametric methods generally are quick and easy to apply, since they involve extremely simple arithmetic. The theory of nonparametric inference relates to properties of the statistic used in the inductive procedure. Discussion of these properties requires derivation of the random sampling distribution of the pertinent statistic, but this generally involves much less sophisticated mathematics than classical statistics. The test statistic in most cases is a discrete random variable with nonzero probabilities assigned to only a finite number of values, and its exact sampling distribution can often be determined by enumeration or simple combinatorial formulas. The asymptotic distributions are usually normal, chi-square, or other well-known functions. The derivations are easier to understand, especially for non-mathematically trained users of statistics. A cookbook approach to learning techniques is then not necessary, which lessens the danger of misuse of procedures. This advantage also minimizes the opportunities for inappropriate and indiscriminate applications, because the assumptions are so general. When no stringent postulations regarding the basic population are needed, there is little problem of violation of assumptions, with the result that conclusions reached in nonparametric methods usually need not be tempered by many qualifiers. The types of assumptions made in nonparametric statistics are generally easily satisfied, and decisions regarding their legitimacy almost obvious. Besides, in many cases the assumptions are sufficient, but not necessary, for the test's validity. Assumptions regarding the sampling process, usually that it is a random sample, are not relaxed with nonparametric methods, but a careful experimenter can generally adopt sampling techniques which render this problem academic. With so-called "dirty data," most nonparametric techniques are, relatively speaking, much more appropriate than parametric methods. The basic data available need not be actual measurements in many cases; if the test is to be based on ranks, for example, only the ranks are needed. The process of collecting and compiling sample data then may be less expensive and time consuming. Some new types of problems relating to sample-distribution characteristics are soluble with nonparametric tests. The scope of application is also wider because the techniques may be legitimately applied to phenomena for which it is impractical or impossible to obtain quantitative measurements. When information about actual observed sample magnitudes is provided but not used as such in drawing an inference, it might seem that some of the available information is being discarded, for which one usually pays a price in efficiency. This is really not true, however. The information embodied in these actual magnitudes, which is not directly employed in the inference procedure, really relates to the underlying distribution, information which is not relevant for distribution-free tests. On the other hand, if the underlying distribution is known, a classical approach to testing may legitimately be used and so this would not be a situation requiring nonparametric methods. The information of course may be consciously ignored, say for the purpose of speed or simplicity.

This discussion of relative merits has so far been concerned mainly with the application of nonparametric techniques. Performance is certainly a matter of concern to the experimenter, but generalizations about reliability are always difficult because of varying factors like sample size, significance levels or confidence coefficients, evaluation of the importance of speed, simplicity and cost factors, and the nonexistence of a fixed and universally acceptable criterion of good performance. Box and Anderson (1955) state that “to fulfill the needs of the experimenter, statistical criteria should (1) be sensitive to change in the specific factors tested, (2) be insensitive to changes, of a magnitude likely to occur in practice, in extraneous factors.” These properties, usually called *power* and *robustness*, respectively, are generally agreed upon as the primary requirements of good performance in hypothesis testing. Parametric tests are often derived in such a way that the first requirement is satisfied for an assumed specific probability distribution, e.g., using the likelihood-ratio technique of test construction. However, since such tests are, strictly speaking, not even valid unless the assumptions are met, robustness is of great concern in parametric statistics. On the other hand, nonparametric tests are inherently robust because their construction requires only very general assumptions. One would expect some sacrifice in power to result. It is therefore natural to look at robustness as a performance criterion for parametric tests and power for nonparametric tests. How then do we compare analogous tests of the two types?

Power calculations for any test require knowledge of the probability distribution of the test statistic under the alternative, but the alternatives in nonparametric problems are often extremely general. When the requisite assumptions are met, many of the classical parametric tests are known to be most powerful. In those cases where comparison studies have been made, however, nonparametric tests are frequently almost as powerful, especially for small samples, and therefore may be considered more desirable whenever there is any doubt about assumptions. No generalizations can be made for moderate-sized samples. The criterion of asymptotic relative efficiency is theoretically relevant only for very large samples. When the classical tests are known to be robust, comparisons may also be desirable for distributions which deviate somewhat from the exact parametric assumptions. However, with inexact assumptions, calculation of power of classical tests is often difficult except by *Monte Carlo techniques*, and studies of power here have been less extensive. Either type of test may be more reliable, depending on the particular tests compared and type or degree of deviations assumed. The difficulty with all these comparisons is that they can be made only for specific nonnull distribution assumptions, which are closely related to the conditions under which the parametric test is exactly valid and optimal.

Perhaps the chief advantage of nonparametric tests lies in their very generality, and an assessment of their performance under conditions unrestricted by, and different from, the intrinsic postulates in classical tests seems more expedient. A comparison under more nonparametric conditions would seem especially desirable for two or more nonparametric tests which are designed for the same general hypothesis testing situation. Unlike the body of classical techniques, nonparametric techniques frequently offer a selection from interchangeable methods. With such a choice, some judgments of relative merit would be particularly useful. Power comparisons have been made, predominantly among the many tests designed to detect location differences, but again we must add that even with comparisons of nonparametric tests, power can be determined only with fairly specific distribution assumptions. The relative merits of the different tests depend on the conditions imposed. Comprehensive conclusions are thus still impossible for blanket comparisons of very general tests.

In conclusion, the extreme generality of nonparametric techniques and their wide scope of usefulness, while definite advantages in application, are factors which discourage objective criteria, particularly power, as assessments of performance, relative either to each other or to parametric techniques. The comparison studies so frequently published in the literature are certainly interesting, informative, and valuable, but they do not provide the sought-for comprehensive answers under more nonparametric conditions. Perhaps we can even say that specific comparisons are really contrary to the spirit of nonparametric methods. No definitive rules of choice will be provided in this book. The interested reader will find many pertinent articles in all the statistics journals. This book is a compendium of many of the large number of nonparametric techniques which have been proposed for various inference situations.

Before embarking on a systematic treatment of new concepts, some basic notation and definitions must be agreed upon and the groundwork prepared for development. Therefore, the remainder of this chapter will be devoted to an explanation of the notation adopted here and an abbreviated review of some of those definitions and terms from classical inference which are also relevant to the special world of nonparametric inference. A few new concepts and terms will also be introduced which are uniquely useful in nonparametric theory. The general theory of order statistics will be the subject of Chapter 2, since they play a fundamental role in many nonparametric techniques. Quantiles, coverages, and tolerance limits are also introduced here. Starting with Chapter 3, the important nonparametric techniques will be discussed in turn, organized according to the type of inference

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

problem (hypothesis to be tested) in the case of hypotheses not involving statements about parameters, or the type of sampling situation (one sample, two independent samples, etc.) in the case of distribution-free techniques, or whichever seems more pertinent. Chapters 3 and 4 will treat tests of randomness and goodness-of-fit tests, respectively, both nonparametric hypotheses which have no counterpart in classical statistics. Chapter 5 covers distribution-free tests of hypotheses and confidence interval estimates of the value of a population quantile in the case of one sample or paired samples. These procedures are based on order statistics, signs, and signed ranks. When the relevant quantile is the median, these procedures relate to the value of a location parameter and are analogies to the one-sample (paired-sample) tests for the population mean (mean difference) in classical statistics. Rank-order statistics are also introduced here, and we investigate the relationship between ranks and variate values. Chapter 6 introduces the two-sample problem and covers some distribution-free tests for the hypothesis of identical distributions against general alternatives. Chapter 7 is an introduction to a particular form of nonparametric test statistic, called a linear rank statistic, which is especially useful for testing a hypothesis that two independent samples are drawn from identical populations. Those linear rank statistics which are particularly sensitive to differences only in location and only in scale are the subjects of Chapters 8 and 9, respectively. Chapter 10 extends this situation to the hypothesis that k independent samples are drawn from identical populations. Chapters 11 and 12 are concerned with measures of association and tests of independence in bivariate and multivariate sample situations, respectively. For almost all tests the discussion will center on logical justification, null distribution and moments of the test statistic, asymptotic distribution, and other relevant distribution properties. Whenever possible, related methods of interval estimation of parameters are also included. During the course of discussion, only the briefest attention will be paid to relative merits of comparable tests. Chapter 13 presents some theorems relating to calculation of asymptotic relative efficiency, a possible criterion for evaluating large sample performance of nonparametric tests relative to each other or to parametric tests when *certain* assumptions are met. These techniques are then used to evaluate the efficiency of some of the tests covered earlier. Chapter 14 covers some special tests based on count data.

Numerical examples of applications of the most commonly used nonparametric test and estimation procedures are included after the explanation of the theory. These illustrations of the techniques will serve to solidify the reader's understanding of proper uses of nonparametric methods. All of the solutions show the calculations clearly. In addition, many of the solutions are then repeated using one or more statistical computer packages.

Problems are given at the end of each chapter. The theoretical problems serve to amplify or solidify the explanations of theory given in the text. The applied problems give the reader practice in applications of the methods. Answers to selected problems are given at the end of the book.

1.2 FUNDAMENTAL STATISTICAL CONCEPTS

In this section a few of the basic definitions and concepts of classical statistics are reviewed, but only very briefly since the main purpose is to explain notation and terms taken for granted later on. A few of the new fundamentals needed for the development of nonparametric inference will also be introduced here.

BASIC DEFINITIONS

A *sample space* is the set of all possible outcomes of a random experiment.

A *random variable* is a set function whose domain is the elements of a sample space on which a probability function has been defined and whose range is the set of all real numbers. Alternatively, X is a random variable if for every real number x there exists a probability that the value assumed by the random variable does not exceed x, denoted by $P(X \leqslant x)$ or $F_X(x)$, and called the *cumulative distribution function* (cdf) of X.

The customary practice is to denote the random variable by a capital letter like X and the actual value assumed (value observed in the experiment) by the corresponding letter in lowercase, x. This practice will generally be adhered to in this book. However, it is not always possible, strictly speaking, to make such a distinction. Occasional inconsistencies will therefore be unavoidable, but the statistically sophisticated reader is no doubt already accustomed to this type of conventional confusion in notation.

The mathematical properties of any function F_X which is a cdf of a random variable X are as follows:

1. $F_X(x_1) \leqslant F_X(x_2)$ for all $x_1 \leqslant x_2$, so that F_X is nondecreasing.
2. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
3. $F_X(x)$ is continuous from the right, or, symbolically, as $\varepsilon \to 0$ through positive values, $\lim_{\varepsilon \to 0} F_X(x + \varepsilon) = F_X(x)$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

A random variable X is called *continuous* if its cdf is continuous. Every continuous cdf in this book will be assumed differentiable everywhere with the possible exception of a finite number of points. The derivative of the cdf will be denoted by $f_X(x)$, a nonnegative function called the *probability density function* (pdf) of X. Thus when X is continuous,

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \qquad f_X(x) = \frac{d}{dx}F_X(x) = F'_X(x) \geqslant 0$$

and

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

A random variable is called *discrete* if it can take on only a finite or a countably infinite number of values, called mass points. The *probability mass function* (pmf) of a discrete random variable X is defined as

$$f_X(x) = P(X = x) = F_X(x) - \lim_{\varepsilon\to 0} F_X(X - \varepsilon)$$

where $\varepsilon \to 0$ through positive values. For a discrete random variable $f_X(x) \geqslant 0$ and $\sum_{\text{all}\,x} f_X(x) = 1$, where the expression "all x" is to be interpreted as meaning all x at which $F_X(x)$ is not continuous; in other words the summation is over all the mass points. Thus for a discrete random variable there is a nonzero probability for any mass point, whereas the probability that a continuous random variable takes on any specific fixed value is zero.

The term *probability function* (pf) or *probability distribution* will be used to denote either a pdf or a pmf. For notation, capital letters will always be reserved for the cdf, while the corresponding lowercase letter denotes the pf.

The *expected value* of a function $g(X)$ of a random variable X, denoted by $E[g(X)]$, is

$$E[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x)\,dx & \text{if } X \text{ is continuous} \\ \sum_{\text{all}\,x} g(x) f_X(x) & \text{if } X \text{ is discrete} \end{cases}$$

Joint probability functions and expectations for functions of more than one random variable are similarly defined and denoted by replacing single symbols by vectors, sometimes abbreviated to

$$\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

A set of n random variables $(X_1, X_2, \ldots, X_n)$ is *independent* if and only if their joint probability function equals the product of the n individual marginal probability functions.

A set of n random variables $(X_1, X_2, \ldots, X_n)$ is called a *random sample* of the random variable X (or from the population F_X or f_X) if they are independent and identically distributed (i.i.d.) so that their joint probability density function is given by

$$f_{\mathbf{X}_n}(x_1, x_2, \ldots, x_n) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$$

A *statistic* is any function of observable or sample random variables which does not involve unknown parameters.

A *moment* is a particular type of population parameter. The kth moment of X about the origin is $\mu'_k = E(X^k)$, where $\mu'_1 = E(X) = \mu$, is the *mean* and the kth *central moment* about the mean is

$$\mu_k = E(X - \mu)^k$$

The second central moment about the mean μ_2 is the *variance* of X,

$$\mu_2 = \text{var}(X) = \sigma^2(X) = E(X^2) - \mu^2 = \mu'_2 - (\mu'_1)^2$$

The kth *factorial moment* is $E[X(X-1)\cdots(X-k+1)]$.

For two random variables, their *covariance* and *correlation*, respectively, are

$$\text{cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X \mu_Y$$

$$\text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

The *moment-generating* function (mgf) of a function g(X) of X is

$$M_{g(X)}(t) = E\{\exp[tg(X)]\}$$

Some special properties of the mgf are

$$M_{a+bX}(t) = e^{bt} M_X(at) \qquad \text{for } a \text{ and } b \text{ constant}$$

$$\mu'_k = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0} = M_X^{(k)}(0)$$

MOMENTS OF LINEAR COMBINATIONS OF RANDOM VARIABLES

Let $X_1, X_2, \ldots, X_n$ be n random variables and $a_i, b_i, i = 1, 2, \ldots, n$ be any constants. Then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

$$\operatorname{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \operatorname{var}(X_i) + 2 \sum_{1 \leqslant i<j \leqslant n}\sum a_i a_j \operatorname{cov}(X_i, X_j)$$

$$\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} b_i X_i\right)$$
$$= \sum_{i=1}^{n} a_i b_i \operatorname{var}(X_i) + \sum_{1 \leqslant i<j \leqslant n}\sum (a_i b_j + a_j b_i) \operatorname{cov}(X_i, X_j)$$

PROBABILITY FUNCTIONS

Some special probability functions are shown in Table 2.1, along with the corresponding mean, variance, and moment-generating function. Both discrete and continuous distributions are included; for a discrete distribution the probability function means the pmf, whereas for a continuous distribution the probability function stands for the corresponding pdf. The term *standard normal* will designate the particular member of the normal family where $\mu = 0$ and $\sigma = 1$. The symbols $\phi(x)$ and $\Phi(x)$ will be reserved for the standard normal density and cumulative distribution functions, respectively.

Three other important distributions are:

$$\text{Student's } t_\nu\text{:} \quad f_X(x) = \frac{\nu^{-1/2}(1 + x^2/\nu)^{-(\nu+1)/2}}{B(\nu/2,\ 1/2)} \qquad \nu > 0$$

Snedecor's $F(\nu_1, \nu_2)$:

$$f_X(x) = \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1} \frac{(1 + \nu_1 x/\nu_2)^{-(\nu_1+\nu_2)/2}}{B(\nu_1/2, \nu_2/2)} \qquad x > 0; \nu_1,\ \nu_2 > 0$$

Fisher's $z(\nu_1, \nu_2)$:

$$f_X(x) = 2\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} e^{\nu_1 x} \frac{(1 + \nu_1 e^{2x}/\nu_2)^{-(\nu_1+\nu_2)/2}}{B(\nu_1/2, \nu_2/2)} \qquad x > 1; \nu_1, \nu_2 > 0$$

The gamma and beta distributions shown in Table 2.1 each contains a special constant, denoted by $\Gamma(\alpha)$ and $B(\alpha, \beta)$ respectively. The *gamma function*, denoted by $\Gamma(\alpha)$, is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \qquad \text{for} \quad \alpha > 0$$

Table 2.1 Some special probability functions

Name	*Probability function* $f_X(x)$	mgf	$E(X)$	var(X)
Discrete distributions				
Bernoulli	$p^x(1-p)^{1-x}$	$pe^t + 1 - p$	p	$p(1-p)$
	$x = 0, 1$			
	$0 \leqslant p \leqslant 1$			
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$(pe^t + 1 - p)^n$	np	$np(1-p)$
	$x = 0, 1, \ldots, n$			
	$0 \leqslant p \leqslant 1$			
Multinomial	$\frac{N!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$	$(p_1 e^{t_1} + \cdots + p_{k-1} e^{t_{k-1}} + p_k)^N$	$E(X_i) = Np_i$	$Var(X_i) = Np_i(1-p_i)$ $\text{Cov}(X_i, X_j) = -Np_i p_j$
	$x_i = 0, 1, \ldots, N;\ \sum x_i = N,$			
	$0 \leqslant p_i \leqslant 1,\ \sum p_i = 1$			
Hypergeometric	$\frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{x}}$		np	$np(1-p)\frac{(N-n)}{(N-1)}$
	$x = 0, 1, \ldots, n$	*		
	$0 \leqslant p \leqslant 1$			

Table 2.1 (Continued)

Name	*Probability function* $f_X(x)$	mgf	*E(X)*	var(*X*)
Geometric	$(1-p)^{x-1}p$ $x = 1, 2, \ldots$ $0 \leqslant p \leqslant 1$	$\dfrac{pe^t}{1-(1-p)e^t}$	$\dfrac{1}{p}$	$\dfrac{(1-p)}{p^2}$
Uniform on $1, 2, \ldots, N$	$\dfrac{1}{N}$ $x = 1, 2, \ldots, N$	$\sum_{x=1}^{N} \dfrac{e^{tx}}{N}$	$\dfrac{N+1}{2}$	$\dfrac{N^2-1}{12}$
Continuous distributions				
Uniform on (α, β)	$\dfrac{1}{\beta-\alpha}$ $\alpha < x < \beta$	$\dfrac{e^{t\beta}-e^{t\alpha}}{t(\beta-\alpha)}$	$\dfrac{\beta+\alpha}{2}$	$\dfrac{(\beta-\alpha)^2}{12}$
Normal	$\dfrac{e^{-(1/2\sigma^2)(x-\mu)^2}}{\sqrt{2\pi\sigma^2}}$ $-\infty < x, \mu < \infty, \sigma > 0$	$e^{(t\mu+t^2\sigma^2/2)}$	μ	σ^2

Gamma	$\dfrac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$ $0 < x,\ \alpha, \beta < \infty$	$(1-\beta t)^{-\alpha}$ $\beta t < 1$	$\alpha\beta$	$\alpha\beta^2$
Exponential	gamma with $\alpha = 1$	$(1-\beta t)^{-1}$ $\beta t < 1$	β	β^2
Chi-square(ν)	gamma with $\alpha = \nu/2,\ \beta = 2$	$(1-2t)^{-1}$ $2t < 1$	ν	2ν
Weibull	$\alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}$ $0 < \alpha, \beta,\ x < \infty$	$\alpha^{-t/\beta},\ \Gamma(1+t/\beta)$	$\alpha^{-1/\beta},\ \Gamma(1+1/\beta)$	$\alpha^{-2/\beta}[\Gamma(1+2/\beta)$ $-\Gamma^2(1+1/\beta)]$
Beta	$\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ $0 < x < 1, \alpha, \beta > 0$	*	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Laplace (double exponential)	$\dfrac{e^{-\lvert x-\theta\rvert/\phi}}{2\phi}$ $-\infty < x, \theta < \infty, \phi > 0$	*	θ	2ϕ
Logistic	$\dfrac{1}{1+e^{-(x-\theta)/\phi}}$ $-\infty < x, \theta < \infty, \phi > 0$	*	θ	$\dfrac{\phi^2\pi^2}{3}$

*The mgf is omitted here because the expression is too complicated.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and has the properties

$$\Gamma(\alpha) = \begin{cases} (\alpha-1)! & \text{for any positive integer } \alpha \\ (\alpha-1)\Gamma(\alpha-1) & \text{for any positive } \alpha, \\ & \quad \text{not necessarily an integer} \\ \sqrt{\pi} & \text{for } \alpha = \frac{1}{2} \end{cases}$$

For other fractional values of α, the gamma function can be found from special tables. For example, $\Gamma(1/4) = 3.6256$, $\Gamma(1/3) = 2.6789$, and $\Gamma(3/4) = 1.2254$ (Abramowitz and Stegun, 1972, p. 255). The *beta function*, denoted by $B(\alpha, \beta)$, is defined as

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx \qquad \text{for } \alpha > 0, \beta > 0$$

The beta and the gamma functions have the following relationship:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

The gamma and the beta functions can be very helpful in evaluating some complicated integrals. For example, suppose we wish to evaluate the integral $I_1 = \int_0^\infty x^4 e^{-x}\,dx$. We identify I_1 as a gamma function $\Gamma(5)$ with $\alpha = 5$ and then $I_1 = \Gamma(5) = 4! = 24$. Similary, the integral $I_2 = \int_0^1 x^4(1-x)^7\,dx$ is a beta function $B(5, 8)$ with $\alpha = 5$ and $\beta = 8$, and thus using the relationship with the gamma function and simplifying, we easily get $I_2 = 4!7!/12! = 1/3960$.

DISTRIBUTIONS OF FUNCTIONS OF RANDOM VARIABLES USING THE METHOD OF JACOBIANS

Let $X_1, X_2, \ldots, X_n$ be n continuous random variables with joint pdf $f(x_1, x_2, \ldots, x_n)$ which is nonzero for an n-dimensional region S_x. Define the transformation

$$y_1 = u(x_1, x_2, \ldots, x_n),$$
$$y_2 = u(x_1, x_2, \ldots, x_n), \quad \ldots, \quad y_n = u(x_1, x_2, \ldots, x_n)$$

which maps S_x onto S_y, where S_x can be written as the union of a finite number m of disjoint spaces $S_1, S_2, \ldots, S_m$ such that the transformation from S_k onto S_y is one to one, for all $k = 1, 2, \ldots, m$. Then for each k there exists a unique inverse transformation, denoted by

$$x_1 = w_{1k}(y_1, y_2, \ldots, y_n),$$
$$x_2 = w_{2k}(y_1, y_2, \ldots, y_n), \ldots, \quad x_n = w_{nk}(y_1, y_2, \ldots, y_n)$$

Assume that for each of these m sets of inverse transformations, the Jacobian

$$J_k(y_1, y_2, \ldots, y_n) = \frac{\partial(w_{1k}, w_{2k}, \ldots, w_{nk})}{\partial(y_1, y_2, \ldots, y_n)} = \det\left(\frac{\partial w_{ik}}{\partial y_i}\right)$$

exists and is continuous and nonzero in S_y, where $\det(a_{ij})$ denotes the determinant of the $n \times n$ matrix with entry a_{ij} in the ith row and jth column. Then the joint pdf of the n random variables $Y_1, Y_2, \ldots, Y_n$, where $Y_i = u_i(X_1, X_2, \ldots, X_n)$, is

$$f(y_1, y_2, \ldots, y_n) = \sum_{k=1}^{m} |J_k(y_1, y_2, \ldots, y_n)| f[w_{1k}(y_1, y_2, \ldots, y_n), \\ w_{2k}(y_1, y_2, \ldots, y_n), \ldots, w_{nk}(y_1, y_2, \ldots, y_n)]$$

for all $(y_1, y_2, \ldots, y_n) \in S_y$, and zero otherwise. The Jacobian of the inverse transformation is the reciprocal of the Jacobian of the direct transformation,

$$\frac{\partial(w_{1k}, w_{2k}, \ldots, w_{nk})}{\partial(y_1, y_2, \ldots, y_n)} = \left[\frac{\partial(u_1, u_2, \ldots, u_n)}{\partial(x_1, x_2, \ldots, x_n)}\right]^{-1}\Bigg|_{x_i = w_{ik}(y_1, y_2, \ldots, y_n)}$$

or

$$J_k(y_1, y_2, \ldots, y_n) = [J_k(x_1, x_2, \ldots, x_n)]^{-1}$$

Thus the pdf above can also be written as

$$f(y_1, y_2, \ldots, y_n) = \sum_{k=1}^{m} |[J_k(x_1, x_2, \ldots, x_n)]^{-1}| f(x_1, x_2, \ldots, x_n)$$

where the right-hand side is evaluated at $x_i = w_{ik}(y_1, y_2, \ldots, y_n)$ for $i = 1, 2, \ldots, n$. If $m = 1$ so that the transformation from S_x onto S_y is one to one, the subscript k and the summation sign may be dropped. When $m = 1$ and $n = 1$, this reduces to the familiar result

$$f_Y(y) = \left[f_X(x)\left|\frac{dy}{dx}\right|^{-1}\right]\Bigg|_{x = u^{-1}(y)}$$

CHEBYSHEV'S INEQUALITY

Let X be any random variable with mean μ and a finite variance σ^2. Then for every $k > 0$, *Chebyshev's inequality* states that

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$P(|X-\mu| \geqslant k) \leqslant \frac{\sigma^2}{k^2}$$

Note that the finite variance assumption guarantees the existence of the mean μ.

The following result, called the *Central Limit Theorem* (CLT), is one of the most famous in statistics. We state it for the simplest i.i.d. situation.

CENTRAL LIMIT THEOREM

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 > 0$ and let $\bar{X}_n$ be the sample mean. Then for $n \to \infty$, the random variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting distribution that is normal with mean 0 and variance 1.

For a proof of this result, typically done via the moment generating function, the reader is referred to any standard graduate level book on mathematical statistics. In some of the non-i.i.d. situations there are other types of CLTs available. For example, if the X's are independent but not identically distributed, there is a CLT generally attributed to Liapounov. We will not pursue these any further.

POINT AND INTERVAL ESTIMATION

A *point estimate* of a parameter is any single function of random variables whose observed value is used to estimate the true value. Let $\hat{\theta}_n = u(X_1, X_2, \ldots, X_n)$ be a point estimate of a parameter θ. Some desirable properties of $\hat{\theta}_n$ are defined as follows for all θ.

1. *Unbiasedness*: $E(\hat{\theta}_n) = \theta$ for all θ.
2. *Sufficiency*: $f_{X_1, X_2, \ldots, X_n | \hat{\theta}_n}(x_1, x_2, \ldots, x_n | \hat{\theta}_n)$ does not depend on θ, or, equivalently,

   $$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta) = g(\hat{\theta}_n; \theta) H(x_1, x_2, \ldots, x_n)$$

   where $H(x_1, x_2, \ldots, x_n)$ does not depend on θ.
3. *Consistency* (also called stochastic convergence and convergence in probability):

   $$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0 \qquad \text{for every } \varepsilon > 0$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

a. If $\hat{\theta}_n$ is an unbiased estimate of θ and $\lim_{n\to\infty} \operatorname{var}(\hat{\theta}_n) = 0$, then $\hat{\theta}_n$ is a consistent estimate of θ, by Chebyshev's inequality.
b. $\hat{\theta}_n$ is a consistent estimate of θ if the limiting distribution of $\hat{\theta}_n$ is a degenerate distribution with probability 1 at θ.

4. *Minimum mean squared error*: $E[(\hat{\theta}_n - \theta)^2] \leqslant E[((\hat{\theta}_n^* - \theta)^2]$, for any other estimate $\hat{\theta}_n^*$.
5. *Minimum variance unbiased*: $\operatorname{var}(\hat{\theta}_n) \leqslant \operatorname{var}(\hat{\theta}_n^*)$ for any other estimate $\hat{\theta}_n^*$ where both $\hat{\theta}_n$ and $\hat{\theta}_n^*$ are unbiased.

An *interval estimate* of a parameter θ with confidence coefficient $1-\alpha$, or a $100(1-\alpha)$ percent *confidence interval* for θ, is a random interval whose end points U and V are functions of observable random variables (usually sample data) such that the probability statement $P(U < \theta < V) = 1 - \alpha$ is satisfied. The probability $P(U < \theta < V)$ should be interpreted as $P(U < \theta) + P(V > \theta)$ since the confidence limits U and V are random variables (depending on the random sample) and θ is a fixed quantity. In many cases this probability can be expressed in terms of a pivotal statistic and the limits can be obtained via tabulated percentiles of standard probability distributions such as the standard normal or the chi-square. A *pivotal statistic* is a function of a statistic and the parameter of interest such that the distribution of the pivotal statistic is free from the parameter (and is often known or at least derivable). For example, $t = \sqrt{n}(\overline{X} - \mu)/S$ is a pivotal statistic for setting up a confidence interval for the mean μ of a normal population with an unknown standard deviation. The random variable t follows a Student's $t_{(n-1)}$ distribution and is thus free from any unknown parameter. All standard books on mathematical statistics cover the topic of confidence interval estimation.

A useful technique for finding point estimates for parameters which appear as unspecified constants (or as functions of such constants) in a family of probability functions, say $f_X(.;\theta)$, is the method of *maximum likelihood*. The *likelihood function* of a random sample of size n from the population $f_X(.;\theta)$ is the joint probability function of the sample variables regarded as a function of θ, or

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

A *maximum-likelihood estimate* (MLE) of θ is a value $\hat{\theta}$ such that for all θ,

$$L(x_1, x_2, \ldots, x_n; \hat{\theta}) \geqslant L(x_1, x_2, \ldots, x_n; \theta)$$

Subject to certain regularity conditions, MLEs are sufficient and consistent and are asymptotically unbiased, minimum variance, and normally distributed. Here, as elsewhere in this book, the term *asymptotic* is interpreted as meaning large sample sizes. Another useful property is *invariance*, which says that if $g(\theta)$ is a *smooth* function of θ and $\hat{\theta}$ is the MLE of θ, then the MLE of $g(\theta)$ is $g(\hat{\theta})$. If more than one parameter is involved, θ above may be interpreted as a vector.

HYPOTHESIS TESTING

A *statistical hypothesis* is a claim or an assertion about the probability function of one or more random variables or a statement about the populations from which one or more samples are drawn, e.g., its form, shape, or parameter values. A hypothesis is called simple if the statement completely specifies the population. Otherwise it is called *composite*. The *null hypothesis* H_0 is the hypothesis under test. The *alternative hypothesis*, H_1 or H_A, is the conclusion reached if the null hypothesis is rejected.

A *test* of a statistical hypothesis is a rule which enables one to make a decision whether or not H_0 should be rejected on the basis of the observed value of a *test statistic*, which is some function of a set of observable random variables. The probability distribution of the test statistic when H_0 holds is sometimes referred to as the *null distribution* of the test statistic.

A *critical region* or *rejection region* R for a test is that subset of values assumed by the test statistic which, in accordance with the test, leads to rejection of H_0. The *critical values* of a test statistic are the bounds of R. For example, if a test statstic T prescribes rejection of H_0 for $T \leqslant t_\alpha$, then t_α is the critical value and R is written symbolically as

$$T \in R \qquad \text{for } T \leqslant t_\alpha$$

A *type I error* is committed if the null hypothesis is rejected when it is true. A *type II error* is failure to reject a false H_0. For a test statistic T of $H_0\colon \theta \in \omega$ versus $H_1\colon \theta \in \Omega - \omega$, the probabilities of these errors are, respectively,

$$\alpha(\theta) = P(T \in R \mid \theta \in \omega) \quad \text{and} \quad \beta(\theta) = P(T \notin R \mid \theta \in \Omega - \omega)$$

The least upper bound value, or supremum, of $\alpha(\theta)$ for all $\theta \in \omega$ is often called the *size of the test. The significance level* is a preselected nominal bound for $\alpha(\theta)$, which may not be attained if the relevant probability function is discrete. Since this is usually the case in

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

nonparametric hypothesis testing, some confusion might arise if these distinctions were adhered to here. So the symbol α will be used to denote either the size of the test or the significance level or the probability of a type I error, prefaced by the adjective "exact" whenever $\sup_{\theta \in \omega} \alpha(\theta) = \alpha$.

The *power* of a test is the probability that the test statistic will lead to a rejection of H_0, denoted by $\text{Pw}(\theta) = P(T \in R)$. Power is of interest mainly as the probability of a correct decision, and so the power is typically calculated when H_0 if false, or H_1 is true, and then $\text{Pw}(\theta) = P(T \in R \mid \theta \in \Omega - \omega) = 1 - \beta(\theta)$. The power depends on the following four variables:

1. The degree of falseness of H_0, that is, the amount of discrepancy between the assertions stated in H_0 and H_1
2. The size of the test α
3. The number of observable random variables involved in the test statistic, generally the sample size
4. The critical region or rejection region R

The *power function* of a test is the power when all but one of these variables are held constant, usually item 1. For example, we can study the power of a particular test as a function of the parameter θ, for a given sample size and α. Typically, the power function is displayed as a plot or a graph of the values of the parameter θ on the X axis against the corresponding power values of the test on the Y axis. To calculate the power of a test, we need the distribution of the test statistic under the alternative hypothesis. Sometimes such a result is either unavailable or is much too complicated to be derived analytically; then *computer simulations* can be used to estimate the power of a test. To illustrate, suppose we would like to estimate the power of a test for the mean μ of a population with $H_0\colon \mu = 10$. We can generate on the computer a random sample from the normal distribution with mean 10 (and say variance equal to 1) and apply the test at a specified level α. If the null hypothesis is rejected, we call it a success. Now we repeat this process of generating a same size sample from the normal distribution with mean 10 and variance 1, say 1000 times. At the end of these 1000 simulations we find the proportion of successes, i.e., the proportion of times when the test rejects the null hypothesis. This proportion is an empirical estimate of the nominal size of a test which was set a priori. To estimate power over the alternative, for example, we can repeat the same process but with samples from a normal distribution with, say, mean 10.5 and variance 1. The proportion of successes from these simulations gives an empirical estimate of the power (simulated power) of the test for the normal distribution when the mean is 10.5 and so on. The simulation technique is particularly useful when a new test is developed with an analytically complicated null and/or alternative distribution and we would like to learn about the test's performance. The number of successes follows a binomial distribution with $n = 1000$ and $p = \alpha$ and this fact can be used to give an idea about the simulation error associated with the proportion of successes in terms of its standard error, which is $\sqrt{\alpha(1-\alpha)/1000}$.

A test is said to be *most powerful* for a specified alternative hypothesis if no other test of the same size has greater power against the same alternative.

A test is *uniformly most powerful* against a class of alternative hypotheses if it is most powerful with respect to each specific simple alternative hypothesis within the class of alternative hypotheses.

A "good" test statistic is one which is reasonably successful in distinguishing correctly between the conditions as stated in the null and alternative hypotheses. A method of constructing tests which often have good properties is the *likelihood-ratio principle*. A random sample of size n is drawn from the population $f_X(.;\theta)$ with likelihood function $L(x_1, x_2, \ldots, x_n; \theta)$, where θ is to be interpreted as a vector if more than one parameter is involved. Suppose that $f_X(.;\theta)$ is a specified family of functions for every $\theta \in \omega$ and ω is a subset of Ω. The *likelihood-ratio* test of

$$H_0\colon \theta \in \omega \qquad \text{versus} \qquad H_1\colon \theta \in \Omega - \omega$$

has the rejection region

$$T \in R \qquad \text{for } T \leqslant c, 0 \leqslant c \leqslant 1$$

where T is the ratio

$$T = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$$

and $L(\hat{\omega})$ and $L(\hat{\Omega})$ are the maximums of the likelihood function with respect to θ for $\theta \in \omega$ and $\theta \in \Omega$, respectively. For an exact size α test of a simple H_0, the number c which defines R is chosen such that $P(T \leqslant c \mid H_0) = \alpha$. Any monotonic function of T, say $g(T)$, can also be employed for the test statistic as long as the rejection region is stated in terms of the corresponding values of $g(T)$; the natural logarithm is one of the most commonly used $g(.)$ functions. The likelihood-ratio test is always a function of sufficient statistics, and the principle often produces a uniformly most powerful test when such exists. A particularly useful property of T for constructing tests based on large samples is that, subject to certain regularity conditions, the probability distribution of $-2 \ln T$ approaches the chi-square distribution with $k_1 - k_2$ degrees of freedom as $n \to \infty$, where k_1 and k_2 are, respectively, the dimensions of the spaces Ω and $\omega, k_2 < k_1$.

All these concepts should be familiar to the reader, since they are an integral part of any standard introductory probability and inference course. We now turn to a few concepts which are especially important in nonparametric inference.

P VALUE

An alternative approach to hypothesis testing is provided by computing a quantity called the *P value*, sometimes called a *probability value* or the *associated probability* or the *significance probability.* A P value is defined as the probability, when the null hypothesis H_0 is true, of obtaining a sample result as extreme as, or more extreme than (in the direction of the alternative), the observed sample result. This probability can be computed for the observed value of the test statistic or some function of it like the sample estimate of the parameter in the null hypothesis. For example, suppose we are testing H_0: $\mu = 50$ versus H_1: $\mu > 50$ and we observe the sample result for $\overline{X}$ is 52. The P value is computed as $P(\overline{X} \geqslant 52 \mid \mu = 50)$. The appropriate direction here is values of $\overline{X}$ that are greater than or equal to 52, since the alternative is μ greater than 50. It is frequently convenient to simply report the P value and go no further. If a P value is small, this is interpreted as meaning that our sample produced a result that is rather rare under the assumption of the null hypothesis. Since the sample result is a fact, it must be that the null hypothesis statement is inconsistent with the sample outcome. In other words, we should reject the null hypothesis. On the other hand, if a P value is large, the sample result is consistent with the null hypothesis and the null hypothesis is not rejected.

If we want to use the P value to reach a decision about whether H_0 should be rejected, we have to select a value for α. If the P value is less than or equal to α, the decision is to reject H_0; otherwise, the decision is not to reject H_0. The P value is therefore the smallest level of significance for which the null hypothesis would be rejected.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

The P value provides not only a means of making a decision about the null hypothesis, but also some idea about how strong the evidence is against the null hypothesis. For example, suppose data set 1 with test T_1 results in a P value of 0.012, while data set 2 with test T_2 (or T_1) has a P value of 0.045. The evidence against the null hypothesis is much stronger for data set 1 than for data set 2 because the observed sample outcome is much less likely in data set 1.

Most of the tables in the Appendix of this book give exact P values for the nonparametric test statistics with small sample sizes. In some books, tables of critical values are given for selected α values. Since the usual α values, 0.01, 0.05, and the like, are seldom attainable exactly for nonparametric tests with small sample sizes, we prefer reporting P values to selecting a level α. If the asymptotic distribution of a test statistic is used to find a P value, this may be called an asymptotic or approximate P value.

If a test has a two-sided alternative, there is no specific direction for calculating the P value. One approach is simply to report the smaller of the two one-tailed P values, indicating that it is one-tailed. If the distribution is symmetric, it makes sense to double this one-tailed P value, and this is frequently done in practice. This procedure is sometimes used even if the distribution is not symmetric.

Finally, note that the P value can be viewed as a random variable. For example, suppose that the test statistic T has a cdf F under H_0 and a cdf G under a one-sided upper-tailed alternative H_1. The P value is the probability of observing a more extreme value than the present random T, so the P value is just the random variable $P = 1 - F(T)$. For a discussion of various properties and ramifications, the reader is referred to Sackrowitz and Samuel-Cahn (1999) and Donahue (1999).

CONSISTENCY

A test is *consistent* for a specified alternative if the power of the test, when that alternative is true, approaches 1 as the sample size approaches infinity. A test is consistent for a class (or subclass) of alternatives if the power of the test when any member of the class (subclass) of alternatives is true approaches 1 as the sample size approaches infinity.

Consistency is a "good" test criterion relevant to both parametric and nonparametric methods, and all the standard test procedures clearly share this property. However, in nonparametric statistics the alternatives are often extremely general, and a wide selection of tests may be available for any one experimental situation. The consistency criterion provides an objective method of choosing among these tests (or at least eliminating some from consideration) when a less general subclass of alternatives is of major interest to the experimenter. A test which is known to be consistent against a specified subclass is said to be especially sensitive to that type of alternative and can generally be recommended for use when the experimenter wishes particularly to detect differences of the type expressed in that subclass.

Consistency of a test can often be shown by investigating whether or not the test statistic converges in probability to the parameter of interest. An especially useful method of investigating consistency is described as follows. A random sample of size n is drawn from the family $f_X(.;\theta), \theta \in \Omega$. Let T be a test statistic for the general hypothesis $\theta \in \omega$ versus $\theta \in \Omega - \omega$, and let $g(\theta)$ be some function of θ such that

$$g(\theta) = \theta_0 \qquad \text{if } \theta \in \omega$$

and

$$g(\theta) \neq \theta_0 \qquad \text{if } \theta \in \Delta \text{ for } \Delta \subset \Omega - \omega$$

If for all θ we have

$$E(T) = g(\theta) \qquad \text{and} \qquad \lim_{n\to\infty} \text{var}(T) = 0$$

then the size α test with rejection region

$$T \in R \qquad \text{for } |T - \theta| > c_\alpha$$

is consistent for the subclass Δ. Similarly, for one-sided subclass of alternatives where

$$g(\theta) = \theta_0 \qquad \text{if } \theta \in \omega$$

and

$$g(\theta) > \theta_0 \qquad \text{if } \theta \in \Delta \text{ for } \Delta \subset \Omega - \omega$$

the consistent test of size a has rejection region

$$T \in R \quad \text{for } T - \theta_0 > c'_\alpha$$

The results follow directly from Chebyshev's inequality. (For a proof, see Fraser, 1957, pp. 267–268.) It may be noted that the unbiasedness condition may be relaxed to asymptotic $(n \to \infty)$ unbiasedness.

PITMAN EFFICIENCY

Another sort of objective criterion may be useful in choosing between two or more tests which are comparable in a well-defined way, namely the concept of *Pitman efficiency*. In the theory of point estimation, the efficiency of two unbiased estimators for a parameter is defined as the ratio of their variances. In some situations, the limiting value of this ratio may be interpreted as the relative number of additional observations needed using the less efficient estimator to obtain the same accuracy. The idea of efficiency of two test statistics is closely related, where power is regarded as a measure of accuracy, but the tests must be compared under equivalent conditions (as both estimators were specified to be unbiased), and there are many variables in hypothesis testing. The most common way to compare two tests is to make all factors equivalent except sample size.

The *power efficiency* of a test A relative to a test B, where both tests are for the same simple null and alternative hypotheses, the same type of rejection region, and the same significance level, is the ratio n_b/n_a, where n_a is the number of observations required by test A for the power of test A to equal the power of test B when n_b observations are employed. Since power efficiency generally depends on the selected significance level, hypotheses, and n_b, it is difficult to calculate and interpret. The problem can be avoided in many cases by defining a type of limiting power efficiency.

Let A and B be two consistent tests of a null hypothesis H_0 and alternative hypothesis H_1, at significance level α. The *asymptotic relative efficiency* (ARE) of test A relative to test B is the limiting value of the ratio n_b/n_a, where n_a is the number of observations required by test A for the power of test A to equal the power of test B based on n_b observations while simultaneously $n_b \to \infty$ and $H_1 \to H_0$.

In many applications of this definition, the ratio is the same for all choices of α, so that the ARE is a single number with a well-defined interpretation for large samples. The requirement that both tests be consistent against H_1 is not a limitation in application, since most tests under consideration for a particular type of alternative will be consistent anyway. But with two consistent tests, their powers both approach 1 with increasing sample sizes. Therefore, we must let H_1 approach H_0 so that the power of each test lies on the open interval $(\alpha, 1)$ for finite sample sizes and the limiting ratio will generally be some number other than 1. The ARE is sometimes also called local asymptotic efficiency since it relates to large sample power in the vicinity of the null hypothesis. A few studies have been conducted which

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

seem to indicate that in several important cases the ARE is a reasonably close approximation to the exact efficiency for moderate-sized samples and alternatives not too far from the null case. Especially in the case of small samples, however, the implications of the ARE value cannot be considered particularly meaningful. Methods of calculating the ARE for comparisons of particular tests will be treated fully in Chapter 13.

The problem of evaluating the relative merits of two or more comparable test statistics is by no means solved by introducing the criteria of consistency and asymptotic relative efficiency. Both are large-sample properties and may not have much import for small- or even moderate-sized samples. As discussed in Section 1.1, exact power calculations are tedious and often too specific to shed much light on the problem as it relates to nonparametric tests, which may explain the general acceptance of asymptotic criteria in the field of nonparametric inference.

The asymptotic relative efficiency of two tests is also defined as the ratio of the limits of the efficacies of the respective tests as the sample sizes approach infinity. The *efficacy* of a test for H_0: $\theta = \theta_0$ based on a sample size n is defined as the derivative of the mean of the test statistic with respect to θ divided by the variance of the test statistic, both evaluated at the hypothesized value $\theta = \theta_0$. Thus, for large n the efficacy measures the rate of change of the mean (expressed in standard units) of a test statistic at the null hypothesis values of θ. A test with a relatively large efficacy is especially sensitive to alternative values of θ close to θ_0 and therefore should have good power in the vicinity of θ_0. Details will be given in Chapter 13.

RANDOMIZED TESTS

We now turn to a different problem which, although not limited to nonparametric inference, is of particular concern in this area. For most classical test procedures, the experimenter chooses a "reasonable" significance level α in advance and determines the rejection-region boundary such that the probability of a type I error is exactly α for a simple hypothesis and does not exceed α for a composite hypothesis. When the null probability distribution of the test statistic is continuous, any real number between 0 and 1 may be chosen as the significance level. Let us call this preselected number the *nominal* α. If the test statistic T can take on only a countable number of values, i.e., if the sampling distribution of T is discrete, the number of possible exact probabilities of a type I error is limited to the number of jump points in the cdf of the test statistic. These exact probabilities will be called *exact* α *values*, or *natural significance levels*. The region can then be chosen such that either (1) the exact α is the largest number which does not exceed the nominal α or (2) the exact α is the number closest to the nominal α. Although most statisticians seem to prefer the first approach, as it is more consistent with classical test procedures for a composite H_0, this has not been universally agreed upon. As a result, two sets of tables of critical values of a test statistic may not be identical for the same nominal α; this can lead to confusion in reading tables. The entries in each table in the Appendix of this book are constructed using the first approach for all critical values.

Disregarding that problem now, suppose we wish to compare the performance, as measured by power, of two different discrete test statistics. Their natural significance levels are unlikely to be the same, so identical nominal α values do not ensure identical exact probabilities of a type I error. Power is certainly affected by exact α, and power comparisons of tests may be quite misleading without identical exact α values. A method of equalizing exact α values is provided by *randomized test procedures*.

A *randomized decision rule* is one which prescribes rejection of H_0 always for a certain range of values of the test statistic, rejection sometimes for another nonoverlapping range, and acceptance otherwise. A typical rejection region of exact size as α might be written $T \in R$ with probability 1 if $T \geqslant t_2$, and with probability p if $t_1 \leqslant T < t_2$, where $t_1 < t_2$ and $0 < p < 1$ are chosen such that

$$P(T \geqslant t_2 | H_0) + pP(t_1 \leqslant T < t_2 | H_0) = \alpha$$

Some random device could be used to make the decision in practice, like drawing one card at random from 100, of which $100p$ are labeled reject. Such decision rules may seem an artificial device and are probably seldom employed by experimenters, but the technique is useful in discussions of theoretical properties of tests. The power of such a randomized test against an alternative H_1 is

$$\text{Pw}(\theta) = P(T \geqslant t_2 | H_1) + pP(t_1 \leqslant T < t_2 | H_1)$$

A simple example will suffice to explain the procedure. A random sample of size 5 is drawn from the Bernoulli population. We wish to test $H_0: \theta = 0.5$ versus $H_1: \theta > 0.5$ at significance level 0.05. The test statistic is X, the number of successes in the sample, which has the

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

binomial distribution with parameter θ and $n = 5$. A reasonable rejection region would be large values of X, and thus the six exact significance levels obtainable without using a randomized test from Table C of the Appendix are:

c	5	4	3	2	1	0
$P(X \geqslant c \mid \theta = 0.5)$	1/32	6/32	16/32	26/32	31/32	1

A nonrandomized test procedure of nominal size 0.05 but exact size

$$\alpha = 1/32 = 0.03125$$

has rejection region

$$X \in R \qquad \text{for } X = 5$$

The randomized test with exact $\alpha = 0.05$ is found with $t_1 = 4$ and $t_2 = 5$ as follows:

$$P(X \geqslant 5 \mid \theta = 0.5) + pP(4 \leqslant X < 5) = 1/32 + pP(X = 4) = 0.05$$

so,

$$1/32 + 5p/32 = 0.05 \text{ and } p = 0.12$$

Thus the rejection region is $X \in R$ with probability 1 if $X = 5$ and with probability 0.12 if $X = 4$. Using Table C, the power of this randomized test when H_1: $\theta = 0.6$ is

$$\begin{aligned}\text{Pw}(0.6) &= P(X = 5 \mid \theta = 0.6) + 0.12\, P(X = 4 \mid \theta = 0.6)\\ &= 0.0778 + 0.12(0.2592) = 0.3110\end{aligned}$$

CONTINUITY CORRECTION

The exact null distribution of most test statistics used in nonparametric inference is discrete. Tables of rejection regions or cumulative distributions are often available for small sample sizes only. However, in many cases some simple approximation to these null distributions is accurate enough for practical applications with moderate-sized samples. When these asymptotic distributions are continuous (like the normal or chi square), the approximation may be improved by introduction a *correction for continuity*. This is accomplished by regarding the value of the discrete test statistic as the midpoint of an interval. For example, if the domain of a test statistic T is only integer values, the observed value is considered to be $t \pm 0.5$. If the decision rule is to reject for $T \geqslant t_{\alpha/2}$ or $T \leqslant t'_{\alpha/2}$ and the large-sample approximation to the distribution of $\frac{T-E(T|H_0)}{\sigma(T|H_0)}$ is the standard normal under H_0, the rejection region with *continuity correction* incorporated is determined by solving the equations

$$\frac{t_{\alpha/2} - 0.5 - E(T|H_0)}{\sigma(T|H_0)} = z_{\alpha/2} \quad \text{and} \quad \frac{t'_{\alpha/2} + 0.5 - E(T|H_0)}{\sigma(T|H_0)} = -z_{\alpha/2}$$

where $z_{\alpha/2}$ satisfies $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. Thus the continuity-corrected, two-sided, approximately size α rejection region is

$$T \geqslant E(T|H_0) + 0.5 + z_{\alpha/2}\sigma(T|H_0) \quad \text{or}$$
$$T \leqslant E(T|H_0) - 0.5 - z_{\alpha/2}\sigma(T|H_0)$$

One-sided rejection regions or critical ratios employing continuity corrections are found similarly. For example, in a one-sided test with rejection region $T \geqslant t_\alpha$, for a nominal size α, the approximation to the rejection region with a continuity correction is determined by solving for t_α in

$$\frac{t_\alpha - 0.5 - E(T|H_0)}{\sigma(T|H_0)} = z_\alpha$$

and thus the continuity corrected, one-sided upper-tailed, approximately size α rejection region is

$$T \geqslant E(T|H_0) + 0.5 + z_\alpha\sigma(T|H_0)$$

Similarly, the continuity corrected, one-sided lower-tailed, approximately size α rejection region is

$$T \leqslant E(T|H_0) - 0.5 - z_\alpha\sigma(T|H_0)$$

The P value for a one-sided test based on a statistic whose null distribution is discrete is often approximated by a continuous distribution, typically the normal, for large sample sizes. Like the rejection regions above, this approximation to the P value can usually be improved by incorporating a correction for continuity. For example, if the alternative is in the upper tail, and the observed value of an integer-valued test statistic T is t_0, the exact P value $P(T \geqslant t_0|H_0)$ is approximated by $P(T \geqslant t_0 - 0.5|H_0)$. In the Bernoulli case with $n = 20$, H_0: $\theta = 0.5$ versus H_1: $\theta > 0.5$, suppose we observe $X = 13$ successes. The normal approximation to the P value with a continuity correction is

$$\begin{aligned} P(X \geqslant 13|H_0) = P(X > 12.5) &= P\left(\frac{X - 10}{\sqrt{5}} > \frac{12.5 - 10}{\sqrt{5}}\right) \\ &= P(Z > 1.12) \\ &= 1 - \Phi(1.12) = 0.1314 \end{aligned}$$

This approximation is very close to the exact P value of 0.1316 from Table C. The approximate P value without the continuity correction is 0.0901, and thus the continuity correction greatly improves the P value approximation. In general, let t_0 be the observed value of the test statistic T whose null distribution can be approximated by the normal distribution. When the alternative is in the upper tail, the approximate P value with a continuity correction is given by

$$1 - \Phi\left[\frac{t_0 - E(T|H_0) - 0.5}{\sigma(T|H_0)}\right]$$

In the lower tail, the continuity corrected approximate P value is given by

$$\Phi\left[\frac{t_0 - E(T|H_0) + 0.5}{\sigma(T|H_0)}\right]$$

When the alternative is two-sided, the continuity corrected approximate P value can be obtained using these two expressions and applying the recommendations given earlier under P value.

2
Order Statistics, Quantiles, and Coverages

2.1 INTRODUCTION

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population with continuous cdf F_X. First let F_X be continuous, so that the probability is zero that any two or more of these random variables have equal magnitudes. In this situation there exists a unique ordered arrangement within the sample. Suppose that $X_{(1)}$ denotes the smallest of the set $X_1, X_2, \ldots, X_n$; $X_{(2)}$ denotes the second smallest; ... and $X_{(n)}$ denotes the largest. Then

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

denotes the original random sample after arrangement in increasing order of magnitude, and these are collectively termed the *order statistics* of the random sample $X_1, X_2, \ldots, X_n$. The rth smallest, $1 \leqslant r \leqslant n$, of the ordered X's, $X_{(r)}$, is called the rth-order statistic. Some familiar

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

applications of order statistics, which are obvious on reflection, are as follows:

1. $X_{(n)}$, the maximum (largest) value in the sample, is of interest in the study of floods and other extreme meteorological phenomena.
2. $X_{(1)}$, the minimum (smallest) value, is useful for phenomena where, for example, the strength of a chain depends on the weakest link.
3. The sample median, defined as $X_{[(n+1)/2]}$ for n odd and any number between $X_{(n/2)}$ and $X_{(n/2+1)}$ for n even, is a measure of location and an estimate of the population central tendency.
4. The sample midrange, defined as $(X_{(1)} + X_{(n)})/2$, is also a measure of central tendency.
5. The sample range $X_{(n)} - X_{(1)}$ is a measure of dispersion.
6. In some experiments, the sampling process ceases after collecting r of the observations. For example, in life-testing electric light bulbs, one may start with a group of n bulbs but stop taking observations after the rth bulb burns out. Then information is available only on the first r ordered "lifetimes" $X_{(1)} < X_{(2)} < \cdots < X_{(r)}$, where $r \leqslant n$. This type of data is often referred to as censored data.
7. Order statistics are used to study outliers or extreme observations, e.g., when so-called dirty data are suspected.

The study of order statistics in this chapter will be limited to their mathematical and statistical properties, including joint and marginal probability distributions, exact moments, asymptotic moments, and asymptotic marginal distributions. Two general uses of order statistics in distribution-free inference will be discussed later in Chapter 5, namely, interval estimation and hypothesis testing of population percentiles. The topic of tolerance limits for distributions, including both one-sample and two-sample coverages, is discussed later in this chapter. But first, we must define another property of probability functions called the quantile function.

2.2 THE QUANTILE FUNCTION

We have already talked about using the mean, the variance, and other moments to describe a probability distribution. In some situations we may be more interested in the percentiles of a distribution, like the fiftieth percentile (the median). For example, if X represents the breaking strength of an item, we might be interested in knowing

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

the median strength, or the strength that is survived by 60 percent of the items, i.e., the fortieth percentile point. Or we may want to know what percentage of the items will survive a pressure of say 3 lb. For questions like these, we need information about the quantiles of a distribution.

A *quantile* of a continuous cdf F_X of a random variable X is a real number that divides the area under the pdf into two parts of specific amounts. Only the area to the left of the number need be specified since the entire area is equal to one. The pth quantile (or the 100pth percentile) of F_X is that value of X, say X_p, such that 100p percent of the values of X in the population are less than or equal to X_p, for any positive fraction $p(0 < p < 1)$. In other words, X_p is a parameter of the population that satisfies $P(X \leqslant X_p) = p$, or, in terms of the cdf $F_X(X_p) = p$. If the cdf of X is strictly increasing, the pth quantile is the unique solution to the equation $X_p = F_X^{-1}(p) = Q_X(p)$, say. We call $Q_X(p),\ 0 < p < 1$, the inverse of the cdf, the *quantile function* (qf) of the random variable X.

Consider, for example, a random variable from the exponential distribution with $\beta = 2$. Then Table 2.1 in Chapter 1 indicates that the cdf is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x/2} & x \geqslant 0 \end{cases}$$

Since $1 - e^{-X_p/2} = p$ for $x > 0$, the inverse is $X_p = -2\ \ln(1 - p)$ for $0 < p < 1$, and hence the quantile function is $Q_X(p) = -2\ \ln(1 - p)$. The cdf and the quantile function for this exponential distribution are shown in Figures 2.1 and 2.2, respectively.

Suppose the distribution of the breaking strength random variable X is this exponential with $\beta = 2$. The reader can verify that the fiftieth percentile $Q_X(0.5)$ is 1.3863, and the fortieth percentile $Q_X(0.4)$ is 1.0217. The proportion that exceeds a breaking strength of 3 pounds is 0.2231.

In general, we define the pth quantile $Q_X(p)$ as the smallest X value at which the cdf is at least equal to p, or

$$Q_X(p) = F_X^{-1}(p) = \inf[x\colon F_X(x) \geqslant p] \qquad 0 < p < 1$$

This definition gives a unique value for the quantile $Q_X(p)$ even when F_X is flat around the specified value p, whereas the previous definition would not give a unique inverse of F_X at p.

Some popular quantiles of a distribution are known as the *quartiles*. The first quartile is the 0.25th quantile, the second quartile

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

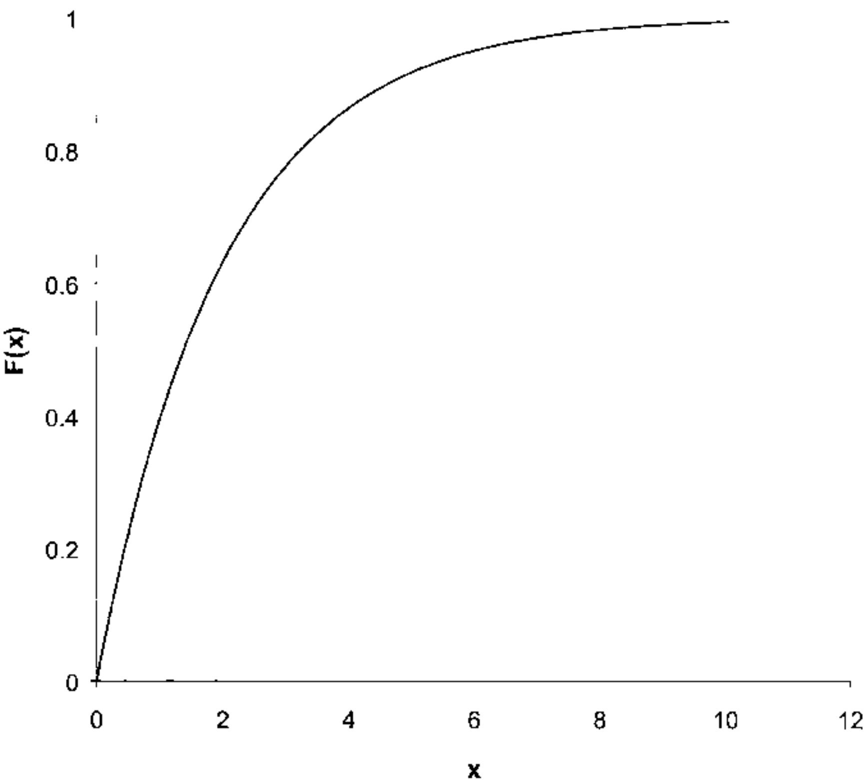


Fig. 2.1 The exponential cdf with $\beta = 2$.

is the 0.50th quantile (the median), and the third quartile is the 0.75th quantile. These are also referred to as the 25th, the 50th, and the 75th percentiles, respectively. Extreme quantiles (such as for $p = 0.95, 0.99$, or 0.995) of a distribution are important as critical values for some test statistics; calculating these is important in many applications.

The cdf and the qf provide similar information regarding the distribution; however, there are situations where one is more natural than the other. Note that formulas for the moments of X can also be expressed in terms of the quantile function. For example,

$$E(X) = \int_0^1 Q_X(p)\,dp \qquad \text{and} \qquad E(X^2) = \int_0^1 Q_X^2(p)\,dp \tag{2.1}$$

so that $\sigma^2 = \int_0^1 Q_X^2(p)\,dp - [\int_0^1 Q_X(p)\,dp]^2$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

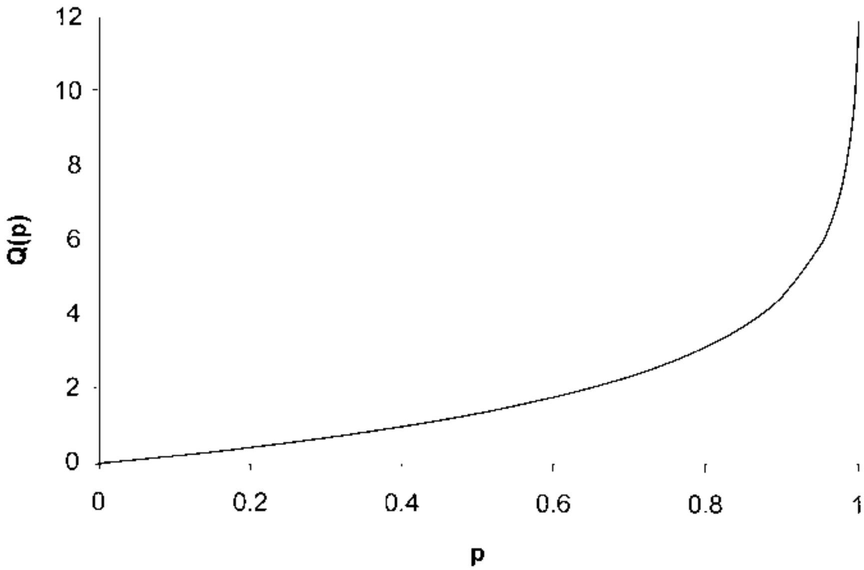


Fig. 2.2 The exponential quantile function with $\beta = 2$.

The following result is useful when working with the qf. Let $f_X(p) = F'_X(p)$ be the pdf of X.

Theorem 2.1 *Assuming that the necessary derivatives all exist, the first and the second derivatives of the quantile function $Q_X(p)$ are*

$$Q'_X(p) = \frac{1}{f_X[Q_X(p)]} \qquad \text{and} \qquad Q''_X(p) = -\frac{f'_X[Q_X(p)]}{\{f_X[Q_X(p)]\}^3}$$

The proof of this result is straightforward and is left for the reader.

It is clear that given some knowledge regarding the distribution of a random variable, one can try to use that information, perhaps along with some data, to aid in studying properties of such a distribution. For example, if we know that the distribution of X is exponential but we are not sure of its mean, typically a simple random sample is taken and the population mean is estimated by the sample mean $\bar{X}$. This estimate can then be used to estimate properties of the distribution. For instance, the probability $P(X \leqslant 3.2)$ can be estimated by $1 - e^{-3.2/\bar{X}}$, which is the estimated cdf of X at 3.2. This, of course, is the approach of classical parametric analysis. In the field of nonparametric analysis, we do not assume that the distribution is exponential (or anything else for that matter). The natural question then is how do we estimate the underlying cdf? This is where the *sample distribution function* (sdf) or the *empirical cumulative distribution function* (ecdf) or the *empirical distribution function* (edf) plays a crucial role.

2.3 THE EMPIRICAL DISTRIBUTION FUNCTION

For a random sample from the distribution F_X, the *empirical distribution function* or edf, denoted by $S_n(x)$, is simply the proportion of sample values less than or equal to the specified value x, that is,

$$S_n(x) = \frac{\text{number of sample values } \leqslant x}{n}$$

In the above example, $S_n(3.2)$ can be used as a point estimate of $P(X \leqslant 3.2)$. The edf is most conveniently defined in terms of the order statistics of a sample, defined in Section 2.1. Suppose that the n sample observations are distinct and arranged in increasing order so that $X_{(1)}$ is the smallest, $X_{(2)}$ is the second smallest, ..., and $X_{(n)}$ is the largest. A formal definition of the edf $S_n(x)$ is

$$S_n(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ i/n & \text{if } X_{(i-1)} \leqslant x < X_{(i)}, i = 1, 2, \ldots, n \\ 1 & \text{if } x \geqslant X_{(n)} \end{cases} \tag{3.1}$$

Suppose that a random sample of size $n = 5$ is given by 9.4, 11.2, 11.4, 12, and 12.6. The edf of this sample is shown in Figure 3.1. Clearly, $S_n(x)$ is a step (or a jump) function, with jumps occuring at the (distinct) ordered sample values, where the height of each jump is equal to the reciprocal of the sample size, namely 1/5 or 0.2.

When more than one observation has the same value, we say these observations are *tied*. In this case the edf is still a step function but it jumps only at the distinct ordered sample values $X_{(j)}$ and the height of the jump is equal to k/n, where k is the number of data values tied at $X_{(j)}$.

We now discuss some of the statistical properties of the edf $S_n(x)$. Let $T_n(x) = nS_n(x)$, so that $T_n(x)$ represents the total number of sample values that are less than or equal to the specified value x.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

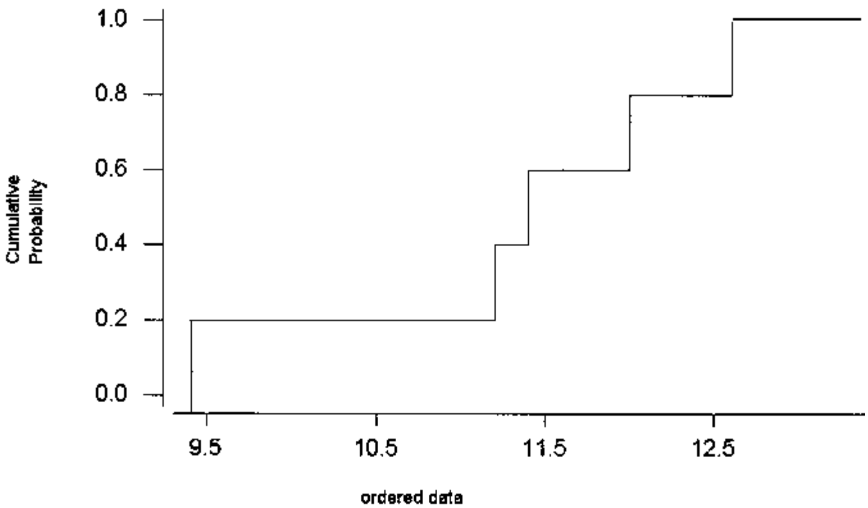


Fig. 3.1 An empirical distribution function for $n = 5$.

Theorem 3.1 *For any fixed real value x, the random variable $T_n(x)$ has a binomial distribution with parameters n and $F_X(x)$.*

Proof For any fixed real constant x and $i = 1, 2, \ldots, n$, define the indicator random variable

$$\delta_i(x) = I_{[X_i \leqslant x]} = \begin{cases} 1 & \text{if } X_i \leqslant x \\ 0 & \text{if } X_i > x \end{cases}$$

The random variables $\delta_1(x), \delta_2(x), \ldots, \delta_n(x)$ are independent and identically distributed, each with the Bernoulli distribution with parameter θ, where $\theta = P[\delta_i(x) = 1] = P(X_i \leqslant x) = F_X(x)$. Now, since $T_n(x) = \sum_{i=1}^{n} \delta_i(x)$ is the sum of n independent and identically distributed Bernoulli random variables, it can be easily shown that $T_n(x)$ has a binomial distribution with parameters n and $\theta = F_X(x)$.

From Theorem 3.1, and using properties of the binomial distribution, we get the following results. The proofs are left for the reader.

Corollary 3.1.1 *The mean and the variance of $S_n(x)$ are*

(a) $E[S_n(x)] = F_X(x)$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

(b) $\text{Var}[S_n(x)] = F_X(x)[1 - F_X(x)]/n$

Part (a) of the corollary shows that $S_n(x)$, the proportion of sample values less than or equal to the specified value x, is an *unbiased* estimator of $F_X(x)$. Part (b) shows that the variance of $S_n(x)$ tends to zero as n tends to infinity. Thus, using Chebyshev's inequality, we can show that $S_n(x)$ is a consistent estimator of $F_X(x)$.

Corollary 3.1.2 *For any fixed real value x, $S_n(x)$ is a* consistent *estimator of $F_X(x)$, or, in other words, $S_n(x)$ converges to $F_X(x)$ in probability.*

Corollary 3.1.3 $E[T_n(x)T_n(y)] = nF_X(x)F_X(y)$, *for $x < y$.*

The convergence in Corollary 3.1.2 is for each value of x individually, whereas sometimes we are interested in all values of x, collectively. A probability statement can be made simultaneously for all x, as a result of the following important theorem. To this end, we have the following classical result [see Fisz (1963), for example, for a proof].

Theorem 3.2 (Glivenko-Cantelli Theorem) *$S_n(x)$ converges uniformly to $F_X(x)$ with probability 1, that is,*

$$P[\lim_{n\to\infty} \sup_{-\infty<x<\infty} |S_n(x) - F_X(x)| = 0] = 1$$

Another useful property of the edf is its asymptotic normality, given in the following theorem.

Theorem 3.3 *As $n \to \infty$, the limiting probability distribution of the standardized $S_n(x)$ is standard normal, or*

$$\lim_{n\to\infty} P\left\{ \frac{\sqrt{n}[S_n(x) - F_X(x)]}{\sqrt{F_X(x)[1 - F_X(x)]}} \leqslant t \right\} = \Phi(t)$$

Proof Using Theorem 3.1, Corollary 3.1.1, and the central limit theorem, it follows that the distribution of $\dfrac{[nS_n(x) - nF_X(x)]}{\sqrt{nF_X(x)[1 - F_X(x)]}} = \dfrac{\sqrt{n}[S_n(x) - F_X(x)]}{\sqrt{F_X(x)[1 - F_X(x)]}}$ approaches the standard normal as $n \to \infty$.

THE EMPIRICAL QUANTILE FUNCTION

Since the population quantile function is the inverse of the cdf and the edf is an estimate of the cdf, it is natural to estimate the quantile function by inverting the edf. This yields the *empirical quantile function* (eqf) $Q_n(u)$, $0 \leqslant u < 1$, defined below.

$$Q_n(u) = \begin{cases} X_{(1)} & \text{if } 0 < u \leqslant 1 \\ X_{(2)} & \text{if } \dfrac{1}{n} < u \leqslant \dfrac{2}{n} \\ X_{(3)} & \text{if } \dfrac{2}{n} < u \leqslant \dfrac{3}{n} \\ \cdots\cdots\cdots\cdots & \\ X_{(n)} & \text{if } \dfrac{n-1}{n} < u \leqslant 1 \end{cases}$$

Thus $Q_n(u) = \inf[x : S_n(x) \geqslant u]$. Accordingly, the empirical (or the sample) quantiles are just the ordered values in a sample. For example, if $n = 10$, the estimate of the 0.30th quantile or the 30th percentile is simply $Q_{10}(0.3) = X_{(3)}$, since $\frac{2}{10} < 0.3 \leqslant \frac{3}{10}$. This is consistent with the usual definition of a quantile or a percentile since 30 percent of the data values are less than or equal to the third order statistic in a sample of size 10. However, note that according to definition, the 0.25th quantile or the 25th percentile (or the 1st quartile) is also equal to $X_{(3)}$ since $2/10 < 0.25 \leqslant 3/10$.

Thus the sample order statistics are point estimates of the corresponding population quantiles. For this reason, a study of the properties of order statistics is as important in nonparametric analysis as the study of the properties of the sample mean in the context of a parametric analysis.

2.4 STATISTICAL PROPERTIES OF ORDER STATISTICS

As we have outlined, the order statistics have many useful applications. In this section we derive some of their statistical properties.

CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF $X_{(r)}$

Theorem 4.1 *For any fixed real t*

$$\begin{aligned} P(X_{(r)} \leqslant t) &= \sum_{i=r}^{n} P[nS_n(t) = i] \\ &= \sum_{i=r}^{n} \binom{n}{i} [F_X(t)]^i [1 - F_X(t)]^{n-i} \qquad -\infty < t < \infty \qquad (4.1) \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

This theorem can be proved in at least two ways. First, $X_{(r)} \leqslant t$ if and only if at least r of the X's are less than or equal to t, and Theorem 3.1 gives the exact distribution of the number of X's less than or equal to t. This result holds even if the underlying distribution is discrete. A second proof, using mathematical statistical results about order statistics, is given later.

PROBABILITY DENSITY FUNCTION (PDF) OF $X_{(r)}$

Theorem 4.2 *If the underlying cdf F_X is continuous with $F'_X(x) = f_X(x)$, the pdf of the* rth-*order statistic is given by*

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!}[F_X(x)]^{r-1}[1-F_X(x)]^{n-r} f_X(x) \qquad -\infty < x < \infty \tag{4.2}$$

This can be proved from Theorem 4.1 by differentiation and some algebraic manipulations. A more direct derivation is provided later.

Theorems 4.1 and 4.2 clearly show that the sample quantiles are not distribution free. Because of this, although intuitively appealing as point estimators of the corresponding population quantiles, these statistics are often not convenient to use except in very special situations. However, they frequently provide interesting starting points and in fact are the building blocks upon which many distribution-free procedures are based. The study of order statistics is thus vital to the understanding of distribution-free inference procedures.

Some important simplification occur when we assume that the sample comes from the continous uniform population on (0,1). Note that for this distribution $F_X(t) = t$ for $0 < t < 1$. Thus, from Theorem 4.1, the cdf of $X_{(r)}$ is

$$F_{X_{(r)}}(t) = P(X_{(r)} \leqslant t) = \sum_{i=r}^{n} P[nS_n(t) = i]$$
$$= \sum_{i=r}^{n} \binom{n}{i} t^i (1-t)^{n-i} \qquad 0 < t < 1$$

and when F is continuous, the pdf of $X_{(r)}$ is a beta distribution given by

$$f_{X_{(r)}}(t) = \frac{n!}{(r-1)!(n-r)!} t^{r-1}(1-t)^{n-r} \qquad 0 < t < 1 \tag{4.3}$$

This is summarized in Theorem 4.3.

Theorem 4.3 *For a random sample of size n from the uniform* (0,1) *distribution, the* rth *order statistic $X_{(r)}$ follows a beta* $(r, n-r+1)$ *distribution.*

The following result follows from Theorems 4.1 and 4.3.

Corollary 4.3.1

$$\sum_{i=r}^{n} \binom{n}{i} t^i (1-t)^{n-i} = \frac{1}{B(r, n-r+1)} \int_0^t x^{r-1}(1-x)^{n-r}\,dx \qquad (4.4)$$

The integral on the right is called an *incomplete beta integral* and is often written as $I_t(r, n-r+1)$. This function has been tabulated by various authors. It can be verified that $1 - I_t(a,b) = I_{1-t}(b,a)$; we leave the verification as an exercise for the reader (Problem 2.3).

2.5 PROBABILITY-INTEGRAL TRANSFORMATION (PIT)

Order statistics are particularly useful in nonparametric statistics because the transformation $U_{(r)} = F(X_{(r)})$ produces a random variable which is the rth-order statistic from the continuous uniform population on the interval (0,1), regardless of what F actually is (normal, gamma, chi-square, etc.); therefore $U_{(r)}$ is distribution free. This property is due to the so-called *probability-integral transformation* (PIT), which is proved in the following theorem.

Theorem 5.1 (Probability-Integral Transformation) *Let X be a random variable with cdf F_X. If F_X is continous, the random variable Y produced by the transformation $Y = F_X(X)$ has the continuous uniform probability distribution over the interval* (0,1).

Proof Since $0 \leqslant F_X(x) \leqslant 1$ for all x, letting F_Y denote the cdf of Y, we have $F_Y(y) = 0$ for $y \leqslant 0$ and $F_Y(y) = 1$ for $y \geqslant 1$. For $0 < y < 1$, define u to be the largest number satisfying $F_X(u) = y$. Then $F_X(X) \leqslant y$ if and only if $X \leqslant u$, and it follows that

$$F_Y(y) = P[F(X) \leqslant y] = P(X \leqslant u) = F_X(u) = y$$

which is the cdf of the continuous uniform distribution defined over (0,1). This completes the proof.

This theorem can also be proved using moment-generating functions when they exist; this approach will be left as an exercise for the reader.

As a result of the PIT, we can conclude that if $X_1, X_2, \ldots, X_n$ is a random sample from any population with continuous distribution F_X, then $F_X(X_1), F_X(X_2), \ldots, F_X(X_n)$ is a random sample from the uniform

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

population. Similarly, if $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the order statistics for the original sample, then

$$F_X(X_{(1)}) < F_X(X_{(2)}) < \cdots < F_X(X_{(n)})$$

are the order statistics from the continuous uniform distribution on (0,1) regardless of the original distribution F_X as long as it is continuous.

The PIT is a very important result in statistics, not only in the theoretical derivations of the properties of order statistics and the like, but also in practical applications such as random number generation. Two examples are now given to illustrate the utility of the PIT.

Example 5.1 Suppose we wish to calculate the probability $P(2 < X \leqslant 3)$, where X follows a chi-square distribution with 3 degrees of freedom (df). Suppose $F_X(X)$ deonotes the cdf of X. Since $F_X(X)$ has the uniform distribution on (0,1) and F_X is nondecreasing, the probability in question is simply equal to $F_X(3) - F_X(2)$. Using the CHIDIST function with df = 3 in the software package EXCEL (note that EXCEL gives right-tail probabilities) we easily get $F_X(2) = 1 - 0.5724 = 0.4276$ and $F_X(3) = 1 - 0.3916 = 0.6084$, so that the required probability is simply $0.6084 - 0.4276 = 0.1808$. Thus transforming the original probability in terms of a probability with respect to the uniform distribution helps simplify the computation.

Example 5.2 An important practical application of the PIT is generating random samples from specified continuous probability distributions. For example, suppose we wish to generate an observation X from an exponential distribution with mean 2. The cdf of X is $F_X(x) = 1 - e^{-x/2}$, and by the PIT, the transformed random variable $Y = 1 - e^{-X/2}$ is distributed as U, an observation from the uniform distribution over the interval (0,1). Now set $1 - e^{-X/2} = U$ and solve for $X = -2\ln(1 - U)$. Using a uniform random number generator (most software packages and some pocket calculators provide one), obtain a uniform random number U and then the desired X from the transformation $X = -2\ln(1 - U)$. Thus, for example, if we get $u = 0.2346$ using a uniform random number generator, the corresponding value of X from the specified exponential distribution is 0.5347.

In order to generate a random sample of 2 or more from a specified continuous probability distribution, we may generate a random sample from the uniform (0,1) distribution and apply the appropriate transformation to each observation in the sample. Several other

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

applications of the probability-integral transformation are given in Problem 2.4.

2.6 JOINT DISTRIBUTION OF ORDER STATISTICS

The joint distribution of order statistics is specified through the joint pdf. Since the observations $X_1, X_2, \ldots, X_n$ in a random sample from a continuous populaiton with pdf f_X are independent and identically distributed random variables, their joint pdf is

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$$

The joint pdf of the n-order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ is not the same as the joint pdf of $X_1, X_2, \ldots, X_n$ since the order statistics are obviously neither independent nor identically distributed. However, the joint pdf is easily derived using the method of Jacobians for transformations.

The set of n order statistics is produced by the transformation

$$\begin{aligned}
Y_1 &= \text{smallest of } (X_1, X_2, \ldots, X_n) = X_{(1)} \\
Y_2 &= \text{second smallest of } (X_1, X_2, \ldots, X_n) = X_{(2)} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
Y_r &= r\text{th smallest of } (X_1, X_2, \ldots, X_n) = X_{(r)} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
Y_n &= \text{largest of } (X_1, X_2, \ldots, X_n) = X_{(n)}
\end{aligned}$$

This transformation is not one to one. In fact, since there are in total $n!$ possible arrangements of the original random variables in increasing order of magnitude, there exist $n!$ inverses to the transformation.

One of these $n!$ permutations might be

$$X_5 < X_1 < X_{n-1} < \cdots < X_n < X_2$$

The corresponding inverse transformation is

$$\begin{aligned}
X_5 &= Y_1 \\
X_1 &= Y_2 \\
X_{n-1} &= Y_3 \\
&\ldots\ldots \\
X_n &= Y_{n-1} \\
X_2 &= Y_n
\end{aligned}$$

The Jacobian of this transformation is the determinant of an $n \times n$ identity matrix with rows rearranged, since each new Y_i is equal to one and only one of the original $X_1, X_2, \ldots, X_n$. The determinant therefore equals ± 1. The joint density function of the random variables in this particular transformation is thus

$$f_{X_1 X_2, \ldots, X_n}(y_2, y_n, \ldots, y_3, y_{n-1})|J| = \prod_{i=1}^{n} f_X(y_i) \quad \text{for } y_1 < y_2 < \cdots < y_n$$

It is easily seen that the same expression results for each of the $n!$ arrangements, since each Jacobian has absolute value 1 and multiplication is commutative. Therefore, applying the general Jacobian technique described in Chapter 1, the result is

$$\begin{aligned} f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(y_1, y_2, \ldots, y_n) &= \sum_{\substack{\text{over all } n! \text{ inverse} \\ \text{transformations}}} \prod_{i=1}^{n} f_X(y_i) \\ &= n! \prod_{i=1}^{n} f_X(y_i) \quad \text{for } y_1 < y_2 < \cdots < y_n \end{aligned} \tag{6.1}$$

In other words, the joint pdf of n order statistics is $n!$ times the joint pdf of the original sample. For example, for a random sample of size n from the normal distribution with mean μ and variance σ^2, we have

$$\begin{aligned} &f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(y_1, y_2, \ldots, y_n) \\ &= \frac{n!}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2} \quad \text{for } -\infty < y_1 < y_2 < \cdots < y_n < \infty \end{aligned}$$

The usual method of finding the marginal pdf of any random variable can be applied to the rth order statistic by integrating out the remaining $(n-1)$ variables in the joint pdf in (6.1). For example, for the largest (maximum) element in the sample, $X_{(n)}$, we have

$$\begin{aligned} f_{X_{(n)}}(y_n) &= n! f_X(y_n) \int_{-\infty}^{y_n} \int_{-\infty}^{y_{n-1}} \cdots \int_{-\infty}^{y_3} \int_{-\infty}^{y_2} \prod_{i=1}^{n-1} f_X(y_i)\, dy_i \\ &= n! f_X(y_n) \int_{-\infty}^{y_n} \int_{-\infty}^{y_{n-1}} \cdots \int_{-\infty}^{y_3} [F_X(y_2) f_X(y_2)] \\ &\quad \times \prod_{i=3}^{n-1} f_X(y_i)\, dy_2 \cdots dy_{n-1} \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$
\begin{aligned}
&= n!\,f_X(y_n)\int_{-\infty}^{y_n}\int_{-\infty}^{y_{n-1}}\cdots\int_{-\infty}^{y_4}\frac{[F_X(y_3)]^2}{2(1)}f_X(y_3)\\
&\quad\times\prod_{i=4}^{n-1}f_X(y_i)\,dy_3\cdots dy_{n-1}\\
&\quad\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&= n!f_X(y_n)\frac{[F_X(y_n)]^{n-1}}{(n-1)!}\\
&= n!\,[F_X(y_n)]^{n-1}f_X(y_n) \qquad (6.2)
\end{aligned}
$$

Similary, for the smallest (minimum) element, $X_{(1)}$, we have

$$
\begin{aligned}
f_{X_{(1)}}(y_1) &= n!f_X(y_1)\int_{y_1}^{\infty}\int_{y_2}^{\infty}\cdots\int_{y_{n-2}}^{\infty}\int_{y_{n-1}}^{\infty}\prod_{i=2}^{n}f_X(y_i)\,dy_n dy_{n-1}\cdots dy_3\,dy_2\\
&= n!f_X(y_1)\int_{y_1}^{\infty}\int_{y_2}^{\infty}\cdots\int_{y_{n-2}}^{\infty}[1-F_X(y_{n-1})]f_X(y_{n-1})\\
&\quad\times\prod_{i=2}^{n-2}f_X(y_i)\,dy_{n-1}dy_{n-2}\cdots dy_2\\
&= n!f_X(y_1)\int_{y_1}^{\infty}\int_{y_2}^{\infty}\cdots\int_{y_{n-3}}^{\infty}\frac{[1-F_X(y_{n-2})]^2}{2(1)}f_X(y_{n-2})\\
&\quad\times\prod_{i=2}^{n-3}f_X(y_i)\,dy_{n-2}\cdots dy_2\\
&\quad\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&= n!f_X(y_1)\frac{[1-F_X(y_1)]^{n-1}}{(n-1)!}\\
&= n!\,[1-F_X(y_1)]^{n-1}f_X(y_1) \qquad (6.3)
\end{aligned}
$$

In general, for the rth-order statistic, the order of integration which is easiest to handle would be $\infty > y_n > y_{n-1} > \cdots > y_r$ followed by $-\infty < y_1 < y_2 < \cdots < y_r$, so that we have the following combination of techniques used for $X_{(n)}$ and $X_{(1)}$:

$$
\begin{aligned}
f_{X_{(r)}(y_r)} &= n!\,f_X(y_r)\int_{-\infty}^{y_r}\int_{-\infty}^{y_{r-1}}\cdots\int_{-\infty}^{y_2}\int_{y_r}^{\infty}\int_{y_{r+1}}^{\infty}\cdots\int_{y_{n-1}}^{\infty}\\
&\quad\times\prod_{\substack{i=1\\ i\neq r}}^{n}f_X(y_i)\,dy_n\cdots dy_{r+2}\,dy_{r+1}\,dy_1\cdots dy_{r-1}
\end{aligned}
$$

$$= n! f_X(y_r) \frac{[1 - F_X(y_r)]^{n-r}}{(n-r)!} \int_{-\infty}^{y_r} \int_{-\infty}^{y_{r-1}} \cdots$$

$$\int_{-\infty}^{y_2} \prod_{i=1}^{r-1} f_X(y_i)\, dy_1 \cdots dy_{r-2}\, dy_{r-1}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$= n! f_X(y_r) \frac{[1 - F_X(y_r)]^{n-r}}{(n-r)!} \frac{[F_X(y_r)]^{r-1}}{(r-1)!}$$

$$= \frac{n!}{(r-1)!(n-r)!} [F_X(y_r)]^{r-1} [1 - F_X(y_r)]^{n-r} f_X(y_r) \quad (6.4)$$

It is clear that this method can be applied to find the marginal distribution of any subset of two or more order statistics and it is relatively easy to apply when finding the joint pdf of a set of successive order statistics, such as $X_{(1)}, X_{(2)}, \ldots, X_{(n-2)}$. In this case we simply integrate out $X_{(n-1)}$ and $X_{(n)}$ as

$$\int_{x_{n-2}}^{\infty} \int_{x_{n-1}}^{\infty} f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n)\, dx_{(n)}\, dx_{(n-1)}$$

The approach, although direct, involves tiresome integration.

A much simpler method can be used which appeals to probability theory instead of pure mathematics. The technique will be illustrated first for the single-order statistic $X_{(r)}$. Recall that by definition of a derivative, we have

$$f_{X_{(r)}(x)} = \lim_{h \to 0} \frac{F_{X_{(r)}}(x+h) - F_{X_{(r)}}(x)}{h}$$

$$= \lim_{h \to 0} \frac{P(x < X_{(r)} \leqslant x - h)}{h} \quad (6.5)$$

Suppose that the x axis is divided into the following three disjoint intervals:

$$I_1 = (-\infty, x]$$
$$I_2 = (x, x+h]$$
$$I_3 = (x+h, \infty)$$

The probability that X lies in each of these intervals is

$$p_1 = P(X \in I_1) = F_X(x)$$
$$p_2 = P(X \in I_2) = F_X(x+h) - F_X(x)$$
$$p_3 = P(X \in I_3) = 1 - F_X(x+h)$$

respectively. Now, $X_{(r)}$ is the rth-order statistic of the set $X_1, X_2, \ldots, X_n$ and lies in the interval I_2 if and only if exactly $r-1$ of the original X random variables lie in the interval I_1, exactly $n-r$ of the original X's lie in the interval I_3 and $X_{(r)}$ lies in the interval I_2. Since the original X values are independent and the intervals are disjoint, the multinomial probability distribution with parameters p_1, p_2, and p_3 can be used to evaluate the probability in (6.5). The result is

$$\begin{aligned} f_{X_{(r)}}(x) &= \lim_{h\to 0} \binom{n}{r-1,\ 1,\ n-r} p_1^{r-1} p_2 p_3^{n-r} \\ &= \frac{n!}{(r-1)!(n-r)!} [F_X(x)]^{r-1} \\ &\quad \times \lim_{h\to 0} \left\{ \frac{F_X(x+h) - F_X(x)}{h} [1 - F_X(x+h)]^{n-r} \right\} \\ &= \frac{n!}{(r-1)!(n-r)!} [F_X(x)]^{r-1} f_X(x) [1 - F_X(x)]^{n-r} \end{aligned} \tag{6.6}$$

which agrees with the result previously obtained in (6.4) and Theorem 4.2.

For the joint distribution, let $X_{(r)}$ and $X_{(s)}$ any two-order statistics from the set $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. By the definition of partial derivatives, the joint pdf can be written

$$\begin{aligned} & f_{X_{(r)},X_{(s)}}(x,y) \\ &= \lim_{\substack{h\to 0\\ t\to 0}} \frac{F_{X_{(r)},X_{(s)}}(x+h,y+t) - F_{X_{(r)},X_{(s)}}(x,y+t) - F_{X_{(r)},X_{(s)}}(x+h,y) + F_{X_{(r)},X_{(s)}}(x,y)}{ht} \\ &= \lim_{\substack{h\to 0\\ t\to 0}} \frac{P(x < X_{(r)} \leqslant x+h, X_{(s)} \leqslant y+t) - P(x < X_{(r)} \leqslant x+h, X_{(s)} \leqslant y)}{ht} \\ &= \lim_{\substack{h\to 0\\ t\to 0}} \frac{P(x < X_{(r)} \leqslant x+h, y < X_{(s)} \leqslant y+t)}{ht} \end{aligned} \tag{6.7}$$

For any $x < y$, the x axis can be divided into the following five disjoint intervals with the corresponding probabilities that an original X observation lies in that interval:

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Interval I	$P(X \in I)$
$I_1 = (-\infty, x]$	$p_1 = F_X(x)$
$I_2 = (x, x+h]$	$p_2 = F_X(x+h) - F_X(x)$
$I_3 = (x+h, y]$	$p_3 = F_X(y) - F_X(x+h)$
$I_4 = (y, y+t]$	$p_4 = F_X(y+t) - F_X(y)$
$I_5 = (y+t, \infty)$	$p_5 = 1 - F_X(y+t)$

With this interval separation and assuming without loss of generality that $r < s$, $X_{(r)}$ and $X_{(s)}$ are the rth- and sth-order statistics, respectively, *and* lie in the respective intervals I_2 and I_4 if and only if the n X values are distributed along the x axis in such a way that exactly $r-1$ lie in I_1, 1 in I_2, 1 in I_4, and $n-s$ in I_5 since the one in I_4 is the sth in magnitude, and the remaining $s-r-1$ must therefore lie in I_3. Applying the multinomial probability distribution to these five types of outcomes with the corresponding probabilities, we obtain

$$\binom{n}{r-1, 1, s-r-1, 1, n-s} p_1^{r-1} p_2 p_3^{s-r-1} p_4 p_5^{n-s}$$

Substituting this for the probability in (6.7) gives

$$\begin{aligned} f_{X_{(r)},X_{(s)}}(x, y) &= \binom{n}{r-1, 1, s-r-1, 1, n-s} [F_X(x)]^{r-1} \\ &\quad \times \lim_{\substack{h\to 0\\ t\to 0}} \left\{ \frac{F_X(x+h) - F_X(x)}{h} [F_X(y) - F_X(x+h)]^{s-r-1} \right\} \\ &\quad \times \lim_{\substack{h\to 0\\ t\to 0}} \left\{ \frac{F_X(y+t) - F_X(y)}{t} [1 - F_X(y+t)]^{n-s} \right\} \\ &= \frac{n!}{(r-1)(s-r-1)!(n-s)!} [F_X(x)]^{r-1} [F_X(y) - F_X(x)]^{s-r-1} \\ &\quad \times [1 - F_X(y)]^{n-s} f_X(x) f_X(y) \quad \text{for all } x < y \end{aligned} \tag{6.8}$$

This method could be extended in a similar manner to find the joint distribution of any subset of the n order statistics. In general, for any $k \leqslant n$, to find the joint distribution of k-order statistics, the x axis must be divided into $k + (k-1) + 2 = 2k + 1$ disjoint intervals and the multinomial probability law applied. For example, the joint pdf of $X_{(r_1)}, X_{(r_2)}, \ldots, X_{(r_k)}$, where $1 \leqslant r_1 < r_2 < \cdots < r_k \leqslant n$ and $1 \leqslant k \leqslant n$ is

$$f_{X_{(r_1)},X_{(r_2)},\ldots,X_{(r_k)}}(x_1,x_2,\ldots,x_k)$$
$$= \frac{n!}{(r_1-1)!(r_2-r_1-1)!\cdots(n-r_k)!}[F_X(x_1)]^{r_1-1}$$
$$\times [F_X(x_2)-F_X(x_1)]^{r_2-r_1-1}\cdots[1-F_X(x_k)]^{n-r_k}$$
$$\times f_X(x_1)f_X(x_2)\cdots f_X(x_k) \qquad x_1 < x_2 < \cdots < x_k$$

In distribution-free techniques we are often interested in the case where $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are order statistics from the continuous uniform distribution over the interval (0, 1). Then $F_X(x) = x$ and so the marginal pdf of $X_{(r)}$ and the joint pdf of $X_{(r)}$ and $X_{(s)}$ for $r < s$ are, respectively,

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!}x^{r-1}(1-x)^{n-r} \qquad 0 < x < 1 \tag{6.9}$$

$$f_{X_{(r)},X_{(s)}}(x,y) = \frac{n!}{(r-1)(s-r-1)!(s-1)!}x^{r-1}(y-x)^{s-r-1}(1-y)^{n-s},$$
$$0 < x < y < 1 \tag{6.10}$$

from (6.4) and (6.8).

The density function in (6.9) will be recognized as that of the beta distribution with parameters r and $n-r+1$. Again, this agrees with the result of Theorem 4.3.

2.7 DISTRIBUTIONS OF THE MEDIAN AND RANGE

As indicated in Section 2.1, the median and range of a random sample are measures based on order statistics which are descriptive of the central tendency and dispersion of the population, respectively. Their distributions are easily obtained from the results found in Section 2.6.

DISTRIBUTION OF THE MEDIAN

For n odd, the median of a sample has the pdf of (6.4) with $r = (n+1)/2$. If n is even and a unique value is desired for the sample median U, the usual definition is

$$U = \frac{X_{(n/2)} + X_{[(n+2)/2]}}{2}$$

so that the distribution of U must be derived from the joint density function of these two-order statistics. Letting $n = 2m$, from (6.8) we have for $x < y$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$f_{X_{(m)},X_{(m+1)}}(x,y) = \frac{(2m)!}{[(m-1)!]^2}[F_X(x)]^{m-1}[1-F_X(y)]^{m-1}f_X(x)f_X(y)$$

Making the transformation

$$u = \frac{x+y}{2}$$
$$v = y$$

and using the method of Jacobians, the pdf of the median U for $n = 2m$ is

$$f_U(u) = \frac{(2m)!2}{[(m-1)!]^2}\int_u^\infty [F_X(2u-v)]^{m-1}[1-F_X(v)]^{m-1} \times f_X(2u-v)f_X(v)\,dv \tag{7.1}$$

As an example, consider the uniform distribution over (0,1). The integrand in (7.1) is nonzero for the intersection of the regions

$$0 < 2u - v < 1 \qquad \text{and} \qquad 0 < v < 1$$

The region of integration then is the intersection of the three regions

$$u < v, \qquad \frac{v}{2} < u < \frac{(v+1)}{2}, \quad \text{and} \quad 0 < v < 1$$

which is depicted graphically in Figure 7.1. We see that the limits on the integral in (7.1) must be $u < v < 2u$ for $0 < u < \frac{1}{2}$ and $u < v < 1$ for $\frac{1}{2} < u < 1$. Thus if $m = 2$, say, the pdf of the median of a sample of size 4 is

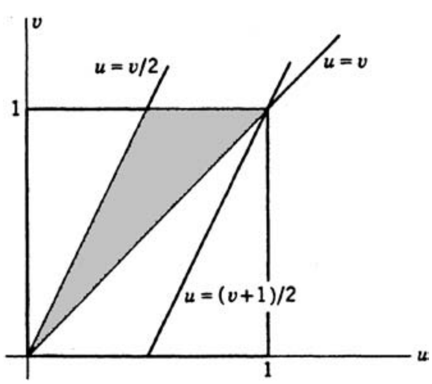


Fig. 7.1 Region of integration is the shaded area.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$f_U(u) = \begin{cases} 8u^2(3-4u) & \text{for } 0 < u \leqslant 1/2 \\ 8(4u^3 - 9u^2 + 6u - 1) & \text{for } 1/2 < u < 1 \end{cases} \tag{7.2}$$

In general, for any integer $m = 1, 2, \ldots$ one can obtain

$$f_U(u) = \begin{cases} \sum_{k=0}^{m-1} \frac{(2m)!2}{k!(m-1)!(m-k-1)!(k+m)} \\ \quad \times (2u-1)^{m-k-1}[(1-u)^{k+m} - (1-2u)^{k+m}] & \text{if } 0 < u \leqslant 1/2 \\ \sum_{k=0}^{m-1} \frac{(2m)!2}{k!(m-1)!(m-k-1)!(k+m)} \\ \quad \times (2u-1)^{m-k-1}(1-u)^{k+m} & \text{if } 1/2 < u < 1 \end{cases}$$

Verification of these results is left for the reader.

DISTRIBUTION OF THE RANGE

A similar procedure can be used to obtain the distribution of the range, defined as

$$R = X_{(n)} - X_{(1)}$$

The joint pdf of $X_{(1)}$ and $X_{(n)}$ is

$$f_{X_{(1)},X_{(n)}}(x, y) = n(n-1)[F_X(y) - F_X(x)]^{n-2} f_X(x) f_X(y) \qquad x < y$$

Now we make the transformation

$$u = y - x$$
$$v = y$$

and obtain, by integration out v, the pdf of the range is

$$f_R(u) = \int_{-\infty}^{\infty} n(n-1)[F_X(v) - F_X(v-u)]^{n-2} f_X(v-u) f_X(v)\, dv \qquad \text{for } u > 0 \tag{7.3}$$

For the uniform distribution, the integrand in (7.3) is nonzero for the intersection of the regions

$$0 < v - u < 1 \qquad \text{and} \qquad 0 < v < 1$$

but this is simply $0 < u < v < 1$. Therefore, the pdf of the range is

$$f_R(u) = n(n-1)u^{n-2}(1-u) \quad \text{for } 0 < u < 1 \tag{7.4}$$

which is the beta distribution with parameters $n-1$ and 2. Thus the result for the uniform distribution is quite easy to handle. However, for a great many distributions, the integral in (7.3) is difficult to evaluate. In the case of a standard normal population, Hartley (1942) has tabulated the cumulative distribution of the range for sample sizes not exceeding 20. The asymptotic distribution of the range is discussed in Gumbel (1944).

2.8 EXACT MOMENTS OF ORDER STATISTICS

Expressions for any individual or joint moments of continuous order statistics can be written down directly using the definition of moments and the specified pdf. The only practical limitation is the complexity of integration involved. Any distribution for which $F_X(x)$ is not easily expressible in a closed form is particularly difficult to handle. In some cases, a more convenient expression for the moments of $X_{(r)}$ can be found in terms of the quantile function $Q_X(u) = F_X^{-1} = F_X^{-1}(u)$ defined in Section 2.2.

*K*TH MOMENT ABOUT THE ORIGIN

The kth moment about the origin of the rth-order statistic from F_X is

$$\begin{aligned} E(X_{(r)}^k) &= \frac{n!}{(r-1)!(n-r)!}\int_{-\infty}^{\infty} y^k[F_X(y)]^{r-1}[1-F_X(y)]^{n-r} f_X(y)\,dy \\ &= \frac{n!}{(r-1)!(n-r)!}\int_{-\infty}^{\infty} y^k[F_X(y)]^{r-1}[1-F_X(y)]^{n-r}\,dF_X(y) \\ &= \frac{n!}{(r-1)!(n-r)!}\int_0^1 [Q_X(u)]^k u^{r-1}(1-u)^{n-r}\,du \\ &= E[Q_X(U)]^k \end{aligned} \tag{8.1}$$

where the random variable U has a beta distribution with parameters r and $n-r+1$. This shows an important relationship between the moments of the order statistics from any arbitrary continuous distribution and the order statistics from the uniform (0,1) distribution. In some cases it may be more convenient to evaluate the integral in (8.1) by numerical methods, especially when a closed-form expression for the quantile function and/or the integral is not readily available.

As an example consider the case of the uniform distribution on the interval (0,1). In this case $Q_X(u) = u$ identically on (0,1) and hence the integral in (8.1) reduces to a beta integral with parameters $r+k$

and $n-r+1$. Thus, using the relationship between the beta and the gamma functions and factorials,

$$\begin{aligned}E(X_{(r)}^k) &= \frac{n!}{(r-1)!(n-r)!}B(r+k,n-r+1)\\ &= \frac{n!}{(r-1)!(n-r)!}\frac{(r+k-1)!(n-r)!}{(n+k)!}\\ &= \frac{n!(r+k-1)!}{(n+k)!(r-1)!}\end{aligned}$$

for any $1 \leqslant r \leqslant n$ and k. In particular, the mean is

$$E(X_{(r)}^k) = \frac{r}{n+1} \tag{8.2}$$

and the variance is

$$\operatorname{var}(X_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)} \tag{8.3}$$

One may immediately recognize (8.2) and (8.3) as the mean and the variance of a beta distribution with parameters r and $n-r+1$. This is of course true since as shown in Theorem 4.3, the distribution of $X_{(r)}$, the rth-order statistic of a random sample of n observations from the uniform (0,1) distribution, is a beta distribution with parameters r and $n-r+1$.

COVARIANCE BETWEEN $X_{(r)}$ AND $X_{(s)}$

Now consider the covariance between any two order statistics $X_{(r)}$ and $X_{(s)}, r<s; r,s=1,2,\ldots,n$, from an arbitrary continuous distribution. From (6.8) we have

$$\begin{aligned}E(X_{(r)}X_{(s)}) = {}& \frac{n!}{(r-1)!(s-r-1)!(n-s)!}\\ &\times \int_{-\infty}^{\infty}\int_{-\infty}^{y} xy[F_X(x)]^{r-1}[F_X(y)-F_X(x)]^{s-r-1}\\ &\times [1-F_X(y)]^{n-s} f_X(x)f_X(y)\,dx\,dy\end{aligned}$$

We now write $f_X(x)\,dx = dF_X(x)$, $f_X(y)\,dy = dF_X(y)$ and substitute $F_X(x)=u$ and $F_X(y)=v$, so that $x = f_X^{-1}(u) = Q_X(u)$ and $y = F_X^{-1}(v) = Q_X(v)$. Then the above expression reduces to

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$\frac{n!}{(r-1)!(s-r-1)!(n-s)!} \times \int_0^1 \int_0^{Q_X(v)} Q_X(u)Q_X(v)u^{r-1}(v-u)^{s-r-1}(1-v)^{n-s}\,du\,dv \qquad (8.4)$$

As remarked before, (8.4) may be more convenient in practice for the actual evaluation of the expectation.

Specializing to the case of the uniform distribution on (0,1) so that $Q_X(u) = u$ and $Q_X(v) = v$, we obtain

$$E(X_{(r)}X_{(s)}) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \times \int_0^1 \int_0^v u^r v(v-u)^{s-r-1}(1-v)^{n-s}\,du\,dv$$

After substituting $z = u/v$ and simplifying, the inner integral reduces to a beta integral and the expectation simplifies to

$$\begin{aligned}
&\frac{n!}{(r-1)!(s-r-1)!(n-s)!} B(r+1,\, s-r) \int_0^1 v^{s+1}(1-v)^{n-s}\,dv \\
&= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} B(r+1,\, s-r) B(s+2, n-s+1) \\
&= \frac{n!r!(s-r-1)!(s+1)!(n-s)!}{(r-1)!(s-r-1)!(n-s)!s!(n+2)!} \\
&= \frac{r(s+1)}{(n+1)(n+2)} \qquad (8.5)
\end{aligned}$$

Now the covariance is found using the formula

$$\text{cov}(X_{(r)}, X_{(s)}) = E(X_{(r)}, X_{(s)}) - E(X_{(r)})E(X_{(s)})$$

which yields, for the uniform (0,1) distribution

$$\begin{aligned}
\text{cov}(X_{(r)}, X_{(s)}) &= \frac{r(s+1)}{(n+1)(n+2)} - \frac{rs}{(n+1)^2} \\
&= \frac{r(n-s+1)}{(n+1)^2(n+2)} \qquad \text{for } r < s \qquad (8.6)
\end{aligned}$$

Thus the correlation coefficient is

$$\text{corr}(X_{(r)}, X_{(s)}) = \left[\frac{(nr-s+1)}{s(n-r+1)}\right]^{1/2} \qquad \text{for } r < s \qquad (8.7)$$

In particular then, the correlation between the minimum and maximum value in a sample of size n from the uniform (0,1) distribution is

$$\operatorname{corr}(X_{(1)}, X_{(n)}) = 1/n$$

which shows that the correlation is inversely proportional to the sample size.

We noted earlier that when the population is such that the cdf $F_X(x)$ or the quantile function $Q_X(u)$ cannot be expressed in a closed form, evaluation of the moments is often tedious or even impossible without the aid of a computer for numerical integration. Since the expected values of the order statistics from a normal probability distribution have especially useful practical applications, these results have been tabulated and are available, for example, in Harter (1961). For small n, these normal moments can be evaluated with appropriate techniques of integration. For example, if $n = 2$ and F_X is the standard normal, the mean of the first-order statistic is

$$\begin{aligned} E(X_{(1)}) &= 2\int_{-\infty}^{\infty} x\left[1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{(-1/2)t^2} dt\right] \frac{1}{\sqrt{2\pi}} e^{(-1/2)x^2}\, dx \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty}\int_{x}^{\infty} x e^{(-1/2)(t^2+x^2)}\, dt\, dx \end{aligned}$$

Introducing a change to polar coordinates with

$$x = r\cos\theta \qquad t = r\sin\theta$$

the integral above becomes

$$\begin{aligned} E(X_{(1)}) &= \frac{1}{\pi}\int_{\pi/4}^{5\pi/4}\int_{0}^{\infty} r^2 \cos\theta e^{(-1/2)r^2}\, dr\, d\theta \\ &= \frac{\sqrt{2}}{\sqrt{\pi}}\int_{\pi/4}^{5\pi/4} \cos\theta\, \frac{1}{2}\int_{-\infty}^{\infty} \frac{r^2}{\sqrt{2\pi}} e^{(-1/2)r^2}\, dr\, d\theta \\ &= \frac{1}{\sqrt{2\pi}}\int_{\pi/4}^{5\pi/4} \cos\theta\, d\theta \\ &= \frac{1}{\sqrt{2\pi}}\left(-\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}\right) = -\frac{1}{\sqrt{\pi}} \end{aligned}$$

Since $E(X_{(1)} + X_{(2)}) = 0$, we have $E(X_{(2)}) = 1/\sqrt{\pi}$.

Other examples of these techniques will be found in the problems.

In nonparametric statistics, we do not assume a parametric form for the underlying population and therefore approximation to the moments of the order statistics is important.

2.9 LARGE-SAMPLE APPROXIMATIONS TO THE MOMENTS OF ORDER STATISTICS

Evaluation of the exact moments of $X_{(r)}$ directly from the pdf could require numerical integration for many F_X of interest. Thus for practical applications and in theoretical investigations, approximations to the moments of $X_{(r)}$ are needed. The PIT plays an important role here since the rth-order statistic from any continuous distribution is a function of the rth-order statistic from the uniform distribution. Letting $U_{(r)}$ denote the rth-order statistic from a uniform distribution over the interval (0,1), this functional relationship can be expressed as

$$X_{(r)} = F_X^{-1}(U_{(r)}) = Q_X(U_{(r)}) \tag{9.1}$$

Now since the moments of $U_{(r)}$ are easily evaluated and $X_{(r)}$ is a function of $U_{(r)}$, the idea is to approximate the moments of $X_{(r)}$ in terms of some function of the moments of $U_{(r)}$. In other words, in studying the moments (or other statistical properties) of $X_{(r)}$ we try to take advantage of the facts that $X_{(r)}$ is a "nice" transformation (function) of the random variable $U_{(r)}$ and that the properties of $U_{(r)}$ are easily found.

Consider first the general case of any random variable Z and any continuous function $g(Z)$ of Z. Since the function $g(Z)$ is continuous, the Taylor series expansion of $g(Z)$ about a point μ is

$$g(Z) = g(\mu) + \sum_{i=1}^{\infty} \frac{(Z-\mu)^i}{i!} g^{(i)}(\mu) \tag{9.2}$$

where $g^{(i)}(\mu) = d^i g(Z)/dZ^i|_{z=\mu}$, and this series converges if

$$\lim_{n\to\infty} \frac{(Z-\mu)^n}{n!} g^{(n)}(z_1) = 0 \qquad \text{for } \mu < z_1 < Z$$

Now if we let $E(Z) = \mu$ and $\operatorname{var}(Z) = \sigma^2$ and take the expectation of both sides of (9.2), we obtain

$$E[g(Z)] = g(\mu) + \frac{\sigma^2}{2!} g^{(2)}(\mu) + \sum_{i=3}^{\infty} \frac{E[(Z-\mu)^i]}{i!} g^{(i)}(\mu) \tag{9.3}$$

From this we immediately see that

1. A first approximation to $E[g(Z)]$ is $g(\mu)$.
2. A second approximation to $E[g(Z)]$ is $g(\mu) + \frac{\sigma^2}{2} g^{(2)}(\mu)$.

To find similar approximations to var(Z), we form the difference between equations (9.2) and (9.3), square this difference, and then take the expectation, as follows:

$$g(Z) - E[g(Z)] = (Z-\mu)g^{(1)}(\mu) + g^{(2)}(\mu)\frac{1}{2!}[(Z-\mu)^2 - \text{var}(Z)] + \sum_{i=3}^{\infty}\frac{g^{(i)}(\mu)}{i!}\{(Z-\mu)^i - E[(Z-\mu)^i]\}$$

$$\{g(Z) - E[g(Z)]\}^2 = (Z-\mu)^2[g^{(1)}(\mu)]^2 + \tfrac{1}{4}[g^{(2)}(\mu)]^2[\text{var}^2(Z) - 2\text{var}(Z)(Z-\mu)^2] - g^{(1)}(\mu)g^{(2)}(\mu)\text{var}(Z)(Z-\mu) + h(Z)$$

so that

$$\text{var}[g(Z)] = \sigma^2[g^{(1)}(\mu)]^2 - \tfrac{1}{4}[g^{(2)}(\mu)]^2\sigma^4 + E[h(Z)] \tag{9.4}$$

where $E[h(Z)]$ involves third or higher central moments of Z.

The first approximations to $E[g(Z)]$ and var$[g(Z)]$ are

$$E[g(Z)] = g(\mu)$$

and

$$\text{var}[g(Z)] = [g^{(1)}(\mu)]^2\sigma^2$$

The second approximations to $E[g(Z)]$ and var$[g(Z)]$ are

$$E[g(Z)] = g(\mu) + \frac{g^{(2)}(\mu)}{2}\sigma^2$$

and

$$\text{var}[g(Z)] = [g^{(1)}(\mu)]^2\sigma^2 - \left[\frac{g^{(2)}(\mu)\sigma^2}{2}\right]^2$$

respectively. The goodness of any of these approximations of course depends on the magnitude of the terms ignored, i.e., the order of the higher central moments of Z.

In order to apply these generally useful results for any random variables to the rth-order statistic of a sample of n from the continuous cdf F_X, we simply take $Z = U_{(r)}$ and note that the functional relationship $X_{(r)} = F_X^{-1}(U_{(r)})$ implies that our g function must be the

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

quantile function, $g(.) = Q_X(.)$. Further, the moments of $U_{(r)}$ were found in (8.2) and (8.3) to be

$$\mu = E(U_{(r)}) = \frac{r}{n+1}$$

and

$$\sigma^2 = \operatorname{var}(U_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)}$$

Also, since the function g is the quantile function given in (9.1), the first two derivatives of the function g, $g^{(1)}$ and $g^{(2)}$, are obtained directly from Theorem 2.1. Evaluating these derivatives at $\mu = r/(n+1)$ we obtain

$$g^{(1)}(\mu) = \left\{ f_X \left[F_X^{-1}\left(\frac{r}{n+1}\right)\right]\right\}^{-1}$$

$$g^{(2)}(\mu) = -f_X^{-1}\left[F_X^{-1}\left(\frac{r}{n+1}\right)\right]\left\{F_X\left[F_X^{-1}\left(\frac{r}{n+1}\right)\right]\right\}^{-3}$$

Substituting these results in the general result above, we can obtain the first and the second approximations to the mean and the variance of $X_{(r)}$. The first approximations are

$$E(X_{(r)}) = F_X^{-1}\left(\frac{r}{n+1}\right) \tag{9.5}$$

and

$$\operatorname{var}(X_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)}\left\{f_X\left[F_x^{-1}\left(\frac{r}{n+1}\right)\right]\right\}^{-2} \tag{9.6}$$

Using (8.1), the third central moment of $U_{(r)}$ can be found to be

$$E[(U_{(r)} - \mu)^3] = \frac{r(2n^2 - 6nr + 4n + 4r^2 - 6r + 2)}{(n+1)^3(n+2)(n+3)} \tag{9.7}$$

so that for large n and finite r or r/n fixed, the terms from (9.3) and (9.4) which were ignored in reaching these approximations are of small order. For greater accuracy, the second- or higher-order approximations can be found. This will be left as an exercise for the reader.

The use of (9.5) and (9.6) is particularly simple when f_X and F_X are tabulated. For example, to approximate the mean and variance of the fourth-order statistic of a sample of 19 from the standard normal population, we have

$$E(X_{(4)}) \approx \Phi^{-1}(0.20) = -0.84$$

$$\text{var}(X_{(4)}) \approx \frac{4(16)}{20^2(21)}[\varphi(-0.84)]^{-2} = \frac{0.16}{21}0.2803^{-2} = 0.097$$

The exact values of the means and variances of the normal order statistics are widely available, for example, in Ruben (1954) and Sarhan and Greenberg (1962). For comparison with the results in this example, the exact mean and variance of $X_{(4)}$ when $n = 19$ are -0.8859 and 0.107406, respectively, from these tables.

2.10 ASYMPTOTIC DISTRIBUTION OF ORDER STATISTICS

As we found in the last section for the moments of order statistics, evaluation of the exact probability density function of $X_{(r)}$ is sometimes rather complicated in practice and it is useful to try to approximate its distribution. When the sample size n is large, such results can be obtained and they are generally called the asymptotic or the large sample distribution of $X_{(r)}$, as n goes to infinity. Information concerning the form of the asymptotic distribution increases the usefulness of order statistics in applications, particularly for large sample sizes. In speaking of a general asymptotic distribution for any r, however, two distinct cases must be considered:

Case 1: As $n \longrightarrow \infty,\ r/n \longrightarrow p,\ 0 < p < 1.$

Case 2: As $n \longrightarrow \infty,\ r$ or $n - r$ remains finite.

Case 1 would be of interest, for example, in the distribution of quantiles, whereas case 2 would be appropriate mainly for the distribution of extreme values. Case 2 will not be considered here. The reader is referred to Wilks (1948) for a discussion of the asymptotic distribution of $X_{(r)}$ for fixed r under various conditions and to Gumbel (1958) for asymptotic distributions of extremes.

Under the assumptions of case 1, we show in this section that the distribution of the standardized rth-order statistic from the uniform distribution approaches the standard normal distribution. This result can be shown in either of two ways. The most direct approach is to show that the probability density function of a standardized $U_{(r)}$ approaches the function $\varphi(u)$. In the density for $U_{(r)}$,

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$f_{U_{(r)}}(u) = \frac{n!}{(r-1)!(n-r)!} u^{r-1}(1-u)^{n-r} \qquad 0 < u < 1$$

we make the transformation

$$Z_{(r)} = \frac{U_{(r)} - u}{\sigma}$$

and obtain, for all z,

$$\begin{aligned} f_{Z_{(r)}}(z) &= \frac{n!}{(r-1)!(n-r)!}(\sigma z + \mu)^{r-1}(1 - \sigma z - \mu)^{n-r}\sigma \\ &= n\binom{n-1}{r-1}\sigma\mu^{r-1}(1-\mu)^{n-r}\left(1 + \frac{\sigma z}{\mu}\right)^{r-1}\left(1 - \frac{\sigma z}{1-\mu}\right)^{n-r} \\ &= n\binom{n-1}{r-1}\sigma\mu^{r-1}(1-\mu)^{n-r}e^{v} \end{aligned} \tag{10.1}$$

where

$$v = (r-1)\ln\left(1 + \frac{\sigma z}{\mu}\right) + (n-r)\ln\left(1 - \frac{\sigma z}{1-\mu}\right) \tag{10.2}$$

Now using the Taylor series expansion

$$\ln(1+x) = \sum_{i=1}^{\infty}(-1)^{i-1}\frac{x^i}{i}$$

which converges for $-1 < x \leqslant 1$, and with the notation

$$\frac{\sigma}{\mu} = c_1 \qquad \frac{\sigma}{1-\mu} = c_2$$

we have

$$\begin{aligned} v &= (r-1)\left(c_1 z - c_1^2\frac{z^2}{2} + c_1^3\frac{z^3}{3} - \cdots\right) - (n-r)\left(c_2 z + c_2^2\frac{z^2}{2} + c_2^3\frac{z^3}{3} + \cdots\right) \\ &= z[c_1(r-1) - c_2(n-r)] - \frac{z^2}{2}[c_1^2(r-1) + c_2^2(n-r)] \\ &\quad + \frac{z^3}{3}[c_1^3(r-1) - c_2^3(n-r)] - \cdots \end{aligned} \tag{10.3}$$

Since we are going to take the limit of v as $n \to \infty, r/n \to p$ fixed, $0 < p < 1$, c_1 and c_2 can be approximated as

$$c_1 = \left[\frac{(n-r+1)}{r(n+2)}\right]^{1/2} \approx \left(\frac{1-p}{pn}\right)^{1/2}$$

$$c_2 = \left[\frac{r}{(n-r+1)(n+2)}\right]^{1/2} \approx \left[\frac{p}{(1-p)n}\right]^{1/2}$$

respectively. Substitution of these values in (10.3) shows that as $n \to \infty$, the coefficient of z is

$$\frac{(r-1)\sqrt{1-p}}{\sqrt{np}} - \frac{(n-r)\sqrt{p}}{\sqrt{n(1-p)}} = \frac{r-np-(1-p)}{\sqrt{np(1-p)}} = -\frac{\sqrt{1-p}}{\sqrt{np}} \to 0$$

the coefficient of $-z^2/2$ is

$$\frac{(r-1)(1-p)}{np} + \frac{(n-r)p}{n(1-p)} = (1-p) - \frac{(1-p)}{np} + p = 1 - \frac{(1-p)}{np} \to 1$$

and the coefficient of $z^3/3$ is

$$\frac{(r-1)(1-p)^{3/2}}{(np)^{3/2}} - \frac{(n-r)p^{3/2}}{[n(1-p)]^{3/2}} = \frac{(np-1)}{n^{3/2}}\left(\frac{1-p}{p}\right)^{3/2} - \frac{p^{3/2}}{[n(1-p)]^{1/2}} \to 0$$

Substituting these results back in (10.3) and ignoring terms of order $n^{-1/2}$ and higher, the limiting value is

$$\lim_{n\to\infty} v = -z^2/2$$

For the limiting value of the constant term in (10.1), we must use *Stirling's formula*

$$k! \approx \sqrt{2\pi}\, e^{-k} k^{k+1/2}$$

for the factorials, which is to be multiplied by

$$\sigma\mu^{r-1}(1-\mu)^{n-r} = \frac{r^{r-1/2}(n-r+1)^{n-r+1/2}}{(n+1)^n(n+2)^{1/2}} \approx \frac{r^{r-1/2}(n-r+1)^{n-r+1/2}}{(n+1)^{n+1/2}}$$

So, as $n \to \infty$, the entire constant of (10.1) is written as

$$\begin{aligned} & n\binom{n-1}{r-1}\sigma\mu^{r-1}(1-\mu)^{n-r} \\ & = \frac{(n+1)!}{r!(n-r+1)!}\frac{r(n-r+1)}{n+1}\sigma\mu^{r-1}(1-\mu)^{n-r} \\ & \approx \frac{\sqrt{2\pi}e^{-(n+1)}(n+1)^{n+3/2}}{2\pi e^{-r}r^{r+1/2}e^{-(n-r+1)}(n-r+1)^{n-r+3/2}}\frac{r^{r+1/2}(n-r+1)^{n-r+3/2}}{(n+1)^{n+3/2}} = \frac{1}{\sqrt{2\pi}} \end{aligned}$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Thus we have the desired result

$$\lim_{n\to\infty} f z_{(r)}(z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2}$$

and hence the *pdf* of the standardized $U_{(r)}$ approaches the *pdf* of the standard normal distribution or, in other words,

$$\lim_{n\to\infty} P(U_{(r)} \leqslant t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

To summarize, for large n, the distribution of $U_{(r)}$ can be approximated by a normal distribution with mean μ and variance σ^2.

For the rth-order statistic from any continuous distribution F_X, the relationship $X_{(r)} = F_X^{-1}(U_{(r)})$ allows us to conclude that the asymptotic distribution of $X_{(r)}$ is also approximately normal as long as the appropriate mean and variance are substituted. The key to this argument is the result that if a random variable is approximately normally distributed then a smooth function of it (a transformation) is also approximately normally distributed with a certain mean and variance. Using the approximate mean and variance found in (9.5) and (9.6) and using the fact that $r/n \to p$ as $n \to \infty$, we get

$$E(X_{(r)}) \to F_X^{-1}(p) \text{ and } \operatorname{var}(X_{(r)}) \approx \frac{[p(1-p)][f_x(\mu)]^{-2}}{n}$$

and state the following theorem.

Theorem 10.1 *Let $X_{(r)}$ denote the* rth-*order statistic of a random sample of size n from any continuous cdf F_X. The if $r/n \to p$ as $n \to \infty, 0 < p < 1$, the distribution of $\left[\frac{n}{p(1-p)}\right]^{1/2} f_X(\mu)[X_{(r)} - \mu]$ tends to the standard normal, where $\mu = F_X^{-1}(p)$.*

Using this result it can be shown that $X_{(r)}$ is a consistent estimator of $\mu = F_X^{-1}(p)$ if $r/n \to p$ as $n \to \infty$.

For the asymptotic joint distribution of any two-order statistics $X_{(r)}$ and $X_{(s)}, 1 \leqslant r < s \leqslant n$, Smirnov (1935) obtained a similar result. Let $n \to \infty$ in such a way that $r/n \to p_1$ and $s/n \to p_2, 0 < p_1 < p_2 < 1$, remain fixed. Then $X_{(r)}$ and $X_{(s)}$ are (jointly) asymptotically bivariate normally distributed with means μ_i, variances $p_i(1-p_i)[f_X(\mu_i)]^{-2}/n$, and covariance $p_1(1-p_2)[nf_X(\mu_1)f_X(\mu_2)]^{-2}$, where μ_i satisfies $F_X(\mu_i) = p_i$ for $i = 1, 2$.

2.11 TOLERANCE LIMITS FOR DISTRIBUTIONS AND COVERAGES

An important application of order statistics is in setting *tolerance limits* for distributions. The resulting procedure does not depend in any way on the underlying population as long as the population is continuous. Such a procedure is therefore distribution free.

A *tolerance interval* for a continuous distribution with tolerance coefficient γ is a random interval (given by two endpoints that are random variables) such that the probability is γ that the area between the endpoints of the interval and under the probability density function is at least a certain preassigned value p. In other words, the probability is γ that this random interval covers or includes at least a specified percentage $(100p)$ of the underlying distribution. If the endpoints of the tolerance interval are two-order statistics $X_{(r)}$ and $X_{(s)}, r < s$, of a random sample of size n, the tolerance interval satisfies the condition

$$P[X_{(r)} < X < X_{(s)} \geqslant p] = \gamma \tag{11.1}$$

The probability γ is called the *tolerance coefficient*. We need to find the two indices r and s, for a given tolerance coefficient, subject to the conditions that $1 \leqslant r < s \leqslant n$. If the underlying distribution F_X is continuous, we can write

$$\begin{aligned} P[X_{(r)} < X < X_{(s)}] &= P(X < X_{(s)}) - P(X < X_{(r)}) \\ &= F_X(X_{(s)}) - F_X(X_{(r)}) \\ &= U_{(s)} - U_{(r)} \end{aligned}$$

according to the PIT. Substituting this result in (11.1), we find that the tolerance interval satisfies

$$P[U_{(s)} - U_{(r)} \geqslant p] = \gamma \tag{11.2}$$

Thus, the question of finding the indices r and s, for any arbitrary continuous distribution reduces to that of finding the indices for the uniform (0,1) distribution. This is a matter of great simplicity, as we show in Theorem 11.1.

Theorem 11.1 *For a random sample of size n from the uniform* (0,1) *distribution, the difference* $U_{(s)} - U_{(r)}$, $1 \leqslant r < s \leqslant n$, *is distributed as the* $(s-r)$th-*order statistic* $U_{(s-r)}$ *and thus has a beta distribution with parameters* $s-r$ *and* $n-s-r+1$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Proof We begin with the joint distribution of $U_{(r)}$ and $U_{(s)}$ found in (6.8). To prove the theorem we make the transformation

$$U = U_{(s)} - U_{(r)} \quad \text{and} \quad V = U_{(s)}$$

The joint distribution of U and V is then

$$f_{U,V}(u, v) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}(v-u)^{r-1}u^{s-r-1}(1-v)^{n-s}$$
$$0 < u < v < 1$$

and so

$$f_U(u) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}u^{s-r-1}\int_u^1 (v-u)^{r-1}(1-v)^{n-s}\,dv$$

Under the integral sign, we make the change of variable $v - u = t(1-u)$ and obtain

$$\begin{aligned} f_U(u) &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!}u^{s-r-1}(1-u)^{r-1} \\ &\quad \times \int_0^1 t^{r-1}[(1-u) - t(1-u)]^{n-s}(1-u)\,dt \\ &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!}u^{s-r-1}(1-u)^{n-s+r}B(r, n-s+1) \\ &= \frac{n!}{(s-r-1)!(n-s+1)!}u^{s-r-1}(1-u)^{n-s+r} \qquad 0 < u < 1 \end{aligned} \tag{11.3}$$

This shows that U has a beta distribution with parameters $s - r$ and $n - s + r + 1$, which is also the distribution of $U_{(s-r)}$ by Theorem 4.3. Thus the required result in (11.2) can be written simply as

$$\gamma = P(U \geqslant p) = \int_p^1 \frac{n!}{(s-r-1)!(n-s+r)!}u^{s-r-1}(1-u)^{n-s+r}\,du$$

We can solve this for r and s for any given values of p and γ, or we can find the tolerance coefficient γ for given values of p, r, and s. Note that all of the above results remain valid as long as the underlying *cdf* is continuous so that the PIT can be applied and hence the tolerance interval is distribution free.

Corollary 11.1.1 *$U_{(r)} - U_{(r-1)}$ has a beta distribution with parameters 1 and n.*

ONE-SAMPLE COVERAGES

The difference $F_X(X_{(s)}) - F_X(X_{(r)}) = U_{(s)} - U_{(r)}$ is called the *coverage* of the random interval $(X_{(r)}, X_{(s)})$, or simply an $s - r$ *cover*. The coverages are generally important in nonparametric statistics because of their distribution-free property. We define the set of successive elementary coverages as the differences.

$$C_i = F_X(X_{(i)}) - F_X(X_{(i-1)}) = U_{(i)} - U_{(i-1)} \qquad i = 1, 2, \ldots, n+1$$

where we write $X_{(0)} = -\infty, X_{(n+1)} = \infty$, Thus,

$$\begin{aligned} C_1 &= F_X(X_{(1)}) = U_{(1)} \\ C_2 &= F_X(X_{(2)}) - F_X(X_{(1)}) = U_{(2)} - U_{(1)} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ C_n &= F_X(X_{(n)}) - F_X(X_{(n-1)}) = U_{(n)} - U_{(n-1)} \\ C_{(n+1)} &= 1 - U_{(n)}) \end{aligned} \tag{11.4}$$

Corollary 11.1.1 shows that the distribution of the ith elementary coverage C_i does not depend on the underlying cdf F_X, as long as F_X is continuous and thus the elementary coverages are distribution-free. In fact, from Corollary 11.1.1 and properties of the beta distribution (or directly), it immediately follows that

$$E(C_i) = \frac{1}{n+1}$$

From this result, we can draw the interpretation that the n-order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ partition the area under the pdf into $n+1$ parts, each of which has the same expected proportion of the total probability.

Since the Jacobian of the transformation defined in (11.4) mapping $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ onto $C_{(1)}, C_{(2)}, \ldots, C_{(n)}$ is equal to 1, the joint distribution of the n coverages is

$$f_{C_1, C_2, \ldots, C_n}(c_1, c_2, \ldots, c_n) = n! \quad \text{for } c_i \geqslant 0,$$
$$i = 1, 2, \ldots, n \text{ and } \sum_{i=1}^{n+1} c_i = 1$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

A sum of any r successive elementary coverages is called an *r coverage*. We have the sum $C_i + C_{i+1} + \cdots + C_{i+r} = U_{(i+r)} - U_{(i)}, i + r \leqslant n$. Since the distribution of $C_1, C_2, \ldots, C_n$ is symmetric in $c_1, c_2, \ldots, c_n$, the marginal distribution of the sum of any r of the coverages must be the same for each fixed value of r, in particular equal to that of

$$C_1 + C_2 + \cdots + C_r = U_{(r)}$$

which is given in (6.9). The expected value of an r coverage then is $r/(n+1)$, with the same interpretation as before.

TWO-SAMPLE COVERAGES

Now suppose that a random sample of size m, $X_1, X_2, \ldots, X_m$ is available from a continuous cdf F_X and that a second independent random sample of size n, $Y_1, Y_2, \ldots, Y_n$ is available from another continuous cdf F_Y. Let $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ be the Y-order statistics and let $I_1 = (-\infty, Y_{(1)}]$, $I_2 = (Y_{(1)}, Y_{(2)}], \ldots, I_n = (Y_{(n-1)}, Y_{(n)}]$, $I_{(n+1)} = (Y_{(n)}, \infty)$ denote the $(n+1)$ nonoverlapping *blocks* formed by the n Y-order statistics. The number of X observations belonging to the ith block, I_i, is called the ith *block frequency* and is denoted by B_i, say. Thus there are $(n+1)$ block frequencies $B_1, B_2, \ldots, B_{n+1}$, where $B_{n+1} = m - B_1 - B_2 - \cdots - B_n$. A particularly appealing feature of the block frequencies is their distribution-free property, summarized in Theorem 11.2.

Theorem 11.2 *When $F_X = F_Y$, that is, the underlying distributions are identical, the joint distribution of $B_1, B_2, \ldots, B_{n+1}$ is given by*

$$P(B_1 = b_1, B_2 = b_2, \ldots, B_{n+1} = b_{n+1}) = \frac{1}{\binom{m+n}{n}}$$

$$\text{where } 0 \leqslant b_j \leqslant m \quad \text{and} \quad \sum_{j=1}^{n+1} b_j = m$$

In fact, one can show that when $F_X = F_Y$, the joint distribution of any t of the random variables $B_1, B_2, \ldots, B_{n+1}$, say $B_1^*, B_2^*, \ldots, B_t^*$, is given by

$$P(B_1^* = b_1^*, B_2^* = b_2^*, \ldots, B_t^* = b_t^*) = \frac{\binom{m+n-b_1^*-b_2^*-\cdots-b_t^*}{n-1}}{\binom{m+n}{n}}$$

$$\text{where } 0 \leqslant b_j^* \leqslant m$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For proofs of these and other related results see, for example, Wilks (1962, pp. 442–446).

We will later discuss a number of popular nonparametric tests based on the block frequencies. Some problems involving the block frequencies are given at the end of this chapter.

RANKS, BLOCK FREQUENCIES, AND PLACEMENTS

The ranks of observations play a crucial role in nonparametric statistics. The rank of the ith observation X_i, in a sample of m observations, is equal to the number of observations that are less than or equal to X_i. In other words, using the indicator function,

$$\text{rank}(X_i) = \sum_{j=1}^{m} I(X_j \leqslant X_i) = mS_m(X_i)$$

where $S_m(X_i)$ is the edf of the sample. For the ordered observation $X_{(i)}$, the rank is simply equal to the index i. That is,

$$\text{rank}(X_{(i)}) = \sum_{j=1}^{m} I(X_j \leqslant X_{(i)}) = mS_m(X_{(i)}) = i$$

Thus, ranks of ordered observations in a single sample are similar to an empirical (data-based) version of the one-sample coverages studied earlier. We provide a functional definition of rank later in Chapter 5 and study some of its statistical properties.

When there are two samples, say m X's and n Y's, the rank of an observation is often defined with respect to the combined sample of $(m+n)$ observations, say Z's. In this case the rank of a particular observation can be defined again as the number of observations (X's and Y's) less than or equal to that particular observation. A functional definition of rank in the two sample case is given later in Chapter 7. However, to see the connection with two-sample coverages, let us examine, for example, the rank of $Y_{(j)}$ in the combined sample. Clearly this is equal to the number of X's less than or equal to $Y_{(j)}$ plus j, the number of Y's less than or equal to $Y_{(j)}$, so that

$$\text{rank}(Y_{(j)}) = \sum_{i=1}^{m} I(X_i \leqslant Y_{(j)}) + j$$

However, $\Sigma_{i=1}^{m} I(X_i \leqslant Y_{(j)})$ is simply equal to $r_1 + r_2 + \cdots + r_j$ where r_i is the frequency of the ith block $(Y_{(i-1)}, Y_{(i)}]$, defined under two-sample coverages. Thus we have

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$\text{rank}(Y_{(j)}) = r_1 + r_2 + \cdots + r_j + j$$

and hence the rank of an ordered Y observation in the combined sample is a simple function of the block frequencies. Also let $P_{(j)} = mS_m(Y_{(j)})$ denote the number of X's that are less than or equal to $Y_{(j)}$. The quantity $P_{(j)}$ is called the *placement* of $Y_{(j)}$ among the X observations (Orban and Wolfe, 1982) and has been used in some nonparametric tests. Then, $P_{(j)} = r_j - r_{j-1}$ with $r_0 = 0$ and $\text{rank}(Y_{(j)}) = P_{(j)} + j$. This shows the connection between ranks and placements. More details regarding the properties of placements are given as problems.

2.12 SUMMARY

In this chapter we discussed some mathematical-statistical concepts and properties related to the distribution function and the quantile function of a random variable. These include order statistics, which can be viewed as sample estimates of quantiles or percentiles of the underlying distribution. However, other methods of estimating population quantiles have been considered in the literature, primarily based on linear functions of order statistics. The reader is referred to the summary section in Chapter 5 for more details.

PROBLEMS

2.1. Let X be a nonnegative continuous random variable with cdf F_X. Show that

$$E(X) = \int_0^\infty [1 - F_X(x)]\, dx$$

2.2. Let X be a discrete random variable taking on only positive integer values. Show that

$$E(X) = \sum_{i=1}^{\infty} P(X \geqslant i)$$

2.3. Show that

$$\sum_{x=a}^{n} \binom{n}{x} p^x (1-p)^{n-x} = \frac{1}{B(a,\ n-a+1)} \int_0^p y^{a-1}(1-y)^{n-a}\, dy$$

for any $0 \leqslant p \leqslant 1$. The integral on the right is called an incomplete beta integral and written as $I_p(a,\ n-a+1)$. Thus, if X is a binomial random variable with parameters n and p, the probability that X is less than or equal to a $(a = 0,\ 1, \ldots, n)$ is

$$1 - I_p(a+1,\ n-a) = I_{1-p}(n-a, a+1)$$

2.4. Find the transformation to obtain, from an observation U following a uniform $(0, 1)$ distribution, an observation from each of the following continuous probability distributions:

(a) Exponential distribution with mean 1.

(b) Beta distribution with $\alpha = 2$ and $\beta = 1$. The probability density function is given by

$$f(x) = 2x \qquad 0 < x < 1$$

(c) The logistic distribution defined by the probability density function

$$f(x) = \frac{e^{-(x-\alpha)/\beta}}{\beta[1 + e^{-(x-\alpha)/\beta}]^2} \qquad -\infty < x < \infty, \quad -\infty < \alpha < \infty, 0 < \beta < \infty$$

(d) The double exponential distribution defined by the probability density function

$$f(x) = \frac{1}{2\beta} e^{-(|x-\alpha|)/\beta} \qquad -\infty < x < \infty, \quad -\infty < \alpha < \infty, 0 < \beta < \infty$$

(e) The Cauchy distribution defined by the probability density function

$$f(x) = \frac{\beta}{\pi[\beta^2 + (x-\alpha)^2]} \qquad -\infty < x < \infty, \quad -\infty < \alpha < \infty, 0 < \beta < \infty$$

2.5. Prove the probability-integral transformation (Theorem 5.1) by finding the moment-generating function of the random variable $Y = F_X(X)$, where X has the continuous cumulative distribution F_X and a moment-generating function that exists.

2.6. If X is a continuous random variable with probability density function $f_X(x) = 2(1 - x), 0 < x < 1$, find the transformation $Y = g(X)$ such that the random variable Y has the uniform distribution over (0,2).

2.7. The order statistics for a random sample of size n from a discrete distribution are defined as in the continuous case except that now we have $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$. Suppose a random sample of size 5 is taken with replacement from the discrete distribution $f_X(x) = 1/6$ for $x = 1, 2, \ldots, 6$. Find the probability mass function of $X_{(1)}$, the smallest order statistic.

2.8. A random sample of size 3 is drawn from the population $f_X(x) = \exp[-(x - \theta)]$ for $x > 0$. We wish to find a 95 percent confidence-interval estimate for the parameter θ. Since the maximum-likelihood estimate for θ is $X_{(1)}$, the smallest order statistic, a logical choice for the limits of the confidence interval would be some functions of $X_{(1)}$. If the upper limit is $X_{(1)}$, find the corresponding lower limit $g(X_{(1)})$ such that the confidence coefficient is 0.95.

2.9. For the n-order statistics of a sample from the uniform distribution over $(0, \theta)$, show that the interval $(X_{(n)},\ X_{(n)}/\alpha^{1/n})$ is a 100 $(1 - \alpha)$ percent confidence-interval estimate of the parameter θ.

2.10. Ten points are chosen randomly and independently on the interval (0,1).

(a) Find the probability that the point nearest 1 exceeds 0.90.

(b) Find the number c such that the probability is 0.5 that the point nearest zero will exceed c.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

2.11. Find the expected value of the largest order statistic in a random sample of size 3 from:

(*a*) The exponential distribution $f_X(x) = \exp(-x)$for $x \geqslant 0$
(*b*) The standard normal distribution

2.12. Verify the result given in (7.1) for the distribution of the median of a sample of size $2m$ from the uniform distribution over (0,1) when $m = 2$. Show that this distribution is symmetric about 0.5 by writing (7.1) in the form

$$f_U(u) = 8(0.5 - |u - 0.5|)^2(1 + 4|u - 0.5|) \qquad 0 < u < 1$$

2.13. Find the mean and variance of the median of a random sample of n from the uniform distribution over (0,1):

(*a*) When n is odd
(*b*) When n is even and U is defined as in Section 2.7.

2.14. Find the probability that the range of a random sample of size n from the population $f_X(x) = 2e^{-2x}$ for $x \geqslant 0$ does not exceed 4.

2.15. Find the distribution of the range of a random sample of size n from the exponential distribution $f_X(x) = 4\exp(-4x)$ for $x \geqslant 0$.

2.16. Give an expression similar to (7.3) for the probability density function of the midrange for any continuous distribution and use it to find the density function in the case of a uniform population over (0,1).

2.17. By making the transformation $U = nF_X(X_{(1)})$, $V = n[1 - F_X(X_{(n)})]$ in (6.8) with $r = 1$, $s = n$, for any continuous F_X, show that U and V are independent random variables in the limiting case as $n \to \infty$, so that the two extreme values of a random sample are asymptotically independent.

2.18. Use (9.5) and (9.6) to approximate the mean and variance of:

(*a*) The median of a sample of size $2m + 1$ from a normal distribution with mean μ and variance α^2.

(*b*) The fifth-order statistic of a random sample of size 19 from the exponential distribution $f_X(x) = \exp(-x)$ for $x \geqslant 0$.

2.19. Let $X_{(n)}$ be the largest value in a sample of size n from the population f_x.

(*a*) Show that $\lim_{n\to\infty} P(n^{-1}X_{(n)} \leqslant x) = \exp(-\alpha/\pi x)$ if $f_X(x) = \alpha/[\pi(\alpha^2 + x^2)]$ (Cauchy).

(*b*) Show that $\lim_{n\to\infty} P(n^{-2}X_{(n)} \leqslant x) = \exp(-\alpha\sqrt{2/\pi x})$ if $f_X(x) = (\alpha/\sqrt{2\pi})x^{-3/2}\exp(-\alpha^2/2x)$ for $x \geqslant 0$.

2.20. Let $X_{(r)}$ be the rth-order statistic of a random sample of size n from a continuous distribution F_X.

(*a*) Show that $P(X_{(r)} \leqslant t) = \sum_{k=r}^{n} \binom{n}{k} [F_X(t)]^k [1 - F_X(t)]^{n-k}$.

(*b*) Verify the probability density function of $X_{(r)}$ given in (6.4) by differentiation of the result in (a).

(*c*) By considering $P(X_{(r)} > t/n)$ in the form of (a), find the asymptotic distribution of $X_{(r)}$ for r fixed and $n \to \infty$ if $F_X(x)$ is the uniform distribution over (0,1).

2.21. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be order statistics for a random sample from the exponential distribution $F_X(x) = \exp(-x)$ for $x \geqslant 0$.

(*a*) Show that $X_{(r)}$ and $X_{(s)} - X_{(r)}$ are independent for any $s > r$.
(*b*) Find the distribution of $X_{(r+1)} - X_{(r)}$.

(*c*) Show that $E(X_{(i)}) = \sum_{j=1}^{i} 1/(n+1-j)$.

(*d*) Interpret the significance of these results if the sample arose from a life test on n light bulbs with exponential lifetimes.

2.22. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the order statistics of a sample from a continuous unspecified distribution F_X. Define the n random variables

$$V_i = \frac{F_X(X_{(i)})}{F_X(X_{(i+1)})} \qquad \text{for } 1 \leqslant i \leqslant n-1 \text{ and } V_n = F_X(X_{(n)})$$

(*a*) Find the marginal distribution of $V_r, 1 \leqslant r \leqslant n$.

(*b*) Find the joint distribution of V_r and $F_X(X_{(r-1)})$, $1 \leqslant r \leqslant n-1$, and show that they are independent.

(*c*) Find the joint distribution of $V_1, V_2, \ldots, V_n$.

(*d*) Show that $V_1, V_2, \ldots, V_n$ are independent.

(*e*) Show that $V_1, V_2^2, V_3^3, \ldots, V_n^n$ are independent and identically distributed with the uniform distribution over (0,1).

2.23. Find the probability that the range of a random sample of size 3 from the uniform distribution is less than 0.8.

2.24. Find the expected value of the range of a random sample of size 3 from the uniform distribution.

2.25. Find the variance of the range of a random sample of size 3 from the uniform distribution.

2.26. Let the random variable U denote the proportion of the population lying between the two extreme values of a sample of n from some unspecified continuous population. Find the mean and variance of U.

2.27. Suppose that a random sample of size m, $X_1, X_2, \ldots, X_m$, is available from a continuous cdf F_X and a second independent random sample of size n, $Y_1, Y_2, \ldots, Y_n$, is available from a continuous cdf F_Y. Let S_j be the random variable representing the number of Y blocks $I_1, I_2, \ldots, I_{n+1}$ (defined in the section on two-sample coverages) that contain exactly j observations from the X sample, $j = 0, 1, \ldots, m$.

(*a*) Verify that $S_0 + S_1 + \cdots + S_m = n + 1$ and $S_1 + 2S_2 + \cdots + mS_m = m$.

(*b*) If $F_X = F_Y$, show that the joint distribution of $S_0, S_1, \ldots, S_m$ is given by $\dfrac{(n+1)!}{s_0! s_1! \cdots s_m!} \binom{m+n}{n}^{-1}$

(*c*) In particular show that, if $F_X = F_Y$, the marginal distribution of S_0 is given by $\binom{n+1}{s_0}\binom{m+1}{n-s_0} \Big/ \binom{m+n}{n}$ for $s_0 = n-m+1, n-m+2, \ldots, n$. (Wilks, 1962)

A simple distribution-free test for the equality of F_X and F_Y can be based on S_0, the number of blocks that do not contain any X observation. This is the "empty block" test (Wilks, 1962, pp. 446–452).

2.28. *Exceedance Statistics*. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples from arbitrary continuous cdf's F_X and F_Y, respectively, and let $S_m(x)$ and $S_n(y)$ be the corresponding empirical cdf's. Consider, for example, the quantity $m[1 - S_m(Y_1)]$, which is simply the count of the total number of X's that exceed (or do not precede) Y_1 and may be called an exceedance statistic. Several nonparametric tests proposed in the literature are based on exceedance (or precedence) statistics and these are called exceedance (or precedence) tests. We will study some of these tests later.

Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ be the order statistics of the Y sample. Answer parts (a) through (h) assuming $F_X = F_Y$.

(a) Show that $S_m(Y_i), i = 1, 2, \ldots, n$, is uniformly distributed over the set of points $(0, 1/m, 2/m, \ldots, 1)$.

(b) Show that the distribution of $S_m(Y_{(j)}) - S_m(Y_{(k)}), k < j$, is the same as the distribution of $S_m(Y_{(j-k)})$. (Fligner and Wolfe, 1976)

(c) Show that the distribution of $P_{(i)} = mS_m(Y_{(i)})$ is given by

$$P[P_{(i)} = j] = \frac{\binom{m+n-i-j}{m-j}\binom{i+j-1}{j}}{\binom{m+n}{n}} \qquad j = 0, 1, \ldots, m$$

The quantity $P_{(i)}$ is the count of the number of X's that precede the ith-order statistic in the Y sample and is called the "placement" of $Y_{(i)}$ among the observations in the X sample. Observe that $P_{(i)} = r_1 + \cdots + r_i$, where r_i is the ith block frequency and thus $r_i = P_{(i)} - P_{(i-1)}$.

(d) Show that

$$E(P_{(i)}) = m\frac{i}{n+1} \qquad \text{and} \qquad \text{var}(P_{(i)}) = \frac{i(n-i+1)m(m+n+1)}{(n+1)^2(n+2)}$$

(Orban and Wolfe, 1982)

(e) Let T_1 be the number of X observations exceeding the largest Y observation, that is, $T_1 = m[1 - S_m(Y_{(n)})] = m - P_{(n)}$. Show that

$$P(T_1 = t) = \frac{\binom{m+n-t-1}{m-t}}{\binom{m+n}{m}}$$

(f) Let T_2 be the number of X's preceding (not exceeding) the smallest Y observation; this is, $T_2 = mS_m(Y_{(1)}) = P_{(1)}$. Show that the distribution of $T_3 = T_1 + T_2$ is given by

$$P(T_3 = t) = (t+1)\frac{\binom{m+n-t-2}{m-t}}{\binom{m+n}{m}} \qquad \text{(Rosenbaum, 1954)}$$

(g) Let T_4 be the number of X's in the interval $I = (Y_{(r)}, Y_{(n+1-r)}]$, where $Y_{(r)}$ is the pth sample quantile of the Y's. The interval I is called the interquartile range of the Y's. Note that $T_4 = m[S_m(Y_{(n+1-r)}) - S_m(Y_{(r)})]$. Show that the distribution of T_4 is given by

$$P(T_4 = t) = \frac{\binom{m+2r-t-1}{m-t}\binom{n+t-2r}{t}}{\binom{m+n}{m}} \qquad t = 0, 1, \ldots, m$$

(*h*) Show that

$$E(T_4) = \frac{2m}{n+1} \qquad \text{and} \qquad \text{var}(T_4) = \frac{2m(n-1)(m+n+1)}{(n+1)^2(n+2)}$$

(Hackl and Katzenbeisser, 1984)

The statistics T_3 and T_4 have been proposed as tests for $H_0: F_X = F_Y$ against the alternative that the dispersion of F_X exceeds the dispersion of F_Y.

2.29. Let $S_m(x)$ be the empirical cdf of a random sample of size m from a continuous cdf F_X. Show that for $-\infty < x < y < \infty$,

$$\text{cov}[S_m(x), S_m(y)] = \frac{F_X(x)[1 - F_X(y)]}{m}$$

2.30. Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution $f_X(x) = (2\theta)^{-1}e^{-x/2\theta}, x \geqslant 0, \theta > 0$, and let the ordered X's be denoted by $Y_1 \leqslant Y_2 \leqslant \cdots \leqslant Y_n$. Assume that the underlying experiment is such that Y_1 becomes available first, then Y_2, and so on (for example, in a life-testing study) and that the experiment is terminated as soon as Y is observed for some specified r.

(*a*) Show that the joint probability density function of $Y_1, Y_2, \ldots, Y_r$ is

$$(2\theta)^{-r}\frac{n!}{(n-r)!}\exp - \left[\frac{\sum_{i=1}^{r} y_i + (n-r)y_r}{2\theta}\right] \qquad 0 \leqslant y_1 \leqslant \cdots \leqslant y_r < \infty$$

(*b*) Show that $\theta^{-1}[\sum_{i=1}^{r} Y_i + (n-r)Y_r]$ has a chi-square distribution with $2r$ degrees of freedom.

2.31. A manufacturer wants to market a new brand of heat-resistant tiles which may be used on the space shuttle. A random sample of m of these tiles is put on a test and the heat resistance capacities of the tiles are measured. Let $X_{(1)}$ denote the smallest of these measurements. The manufacturer is interested in finding the probability that in a future test (performed by, say, an independent agency) of a random sample of n of these tiles, at least k $(k = 1, 2, \ldots, n)$ will have a heat resistance capacity exceeding $X_{(1)}$ units. Assume that the heat resistance capacities of these tiles follows a continuous distribution with cdf F.

(*a*) Show that the probability of interest is given by

$$\sum_{r=k}^{n} P(r)$$

where

$$P(r) = \frac{mn!(r+m-1)!}{r!(n+m)!}$$

(*b*) Show that

$$\frac{P(r-1)}{P(r)} = \frac{r}{r+m-1}$$

a relationship that is useful in calculating $P(r)$.

(c) Show that the number of tiles n to be put on a future test such that all of the n measurements exceed $X_{(1)}$ with probability p is given by

$$n = \frac{m(1-p)}{p}$$

2.32. Define the random variable

$$\varepsilon(x) = \begin{cases} 1 & \text{if } x \geqslant 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Show that the random function defined by

$$F_n(x) = \sum_{i=1}^{n} \frac{\varepsilon(x - X_i)}{n}$$

is the empirical distribution function of a sample $X_1, X_2, \ldots, X_n$, by showing that

$$F_n(x) = S_n(x) \text{ for all } x$$

.

2.33. Prove that $\text{cov}[S_n(x),\ S_n(y)] = c[F_X(x),\ F_X(y)]/n$ where

$$c(s,t) = \min(s,t) - st = \begin{cases} s(1-t) & \text{if } s \leqslant t \\ t(1-s) & \text{if } s \geqslant t \end{cases}$$

and $S_n(.)$ is the empirical distribution function of a random sample of size n from the population F_x.

2.34. Let $S_n(x)$ be the empirical distribution function for a random sample of size n from the uniform distribution on (0,1). Define

$$X_n(t) = \sqrt{n}|S_n(t) - t|$$

$$Z_n(t) = (t+1)X_n\left(\frac{t}{t+1}\right) \qquad \text{for all } 0 \leqslant t \leqslant 1$$

Find $E[X_n(t)]$ and $E[Z_n(t)], \text{var}[X_n(t)]$ and $\text{var}[Z_n(t)]$, and conclude that $\text{var}[X_n(t)] \leqslant \text{var}[Z_n(t)]$ for all $0 \leqslant t \leqslant 1$ and all n.

3
Tests of Randomness

3.1 INTRODUCTION

Passing a line of ten persons waiting to buy a ticket at a movie theater on a Saturday afternoon, suppose we observe the arrangement of five males and five females in the line to be M, F, M, F, M, F, M, F, M, F. Would this be considered a random arrangement by gender? Intuitively, the answer is no, since the alternation of the two types of symbols suggests intentional mixing by pairs. This arrangement is an extreme case, as is the configuration M, M, M, M, M, F, F, F, F, F, with intentional clustering. In the less extreme situations, the randomness of an arrangement can be tested statistically using the theory of runs.

Given an ordered sequence of one or more types of symbols, a *run* is defined to be a succession of one or more types of symbols which are followed and preceded by a different symbol or no symbol at all. Clues to lack of randomness are provided by any tendency of the symbols to

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

exhibit a definite pattern in the sequence. Both the number of runs and the lengths of the runs, which are of course interrelated, should reflect the existence of some sort of pattern. Tests for randomness can therefore be based on either criterion or some combination thereof. Too few runs, too many runs, a run of excessive length, too many runs of excessive length, etc. can be used as statistical criteria for rejection of the null hypothesis of randomness, since these situations should occur rarely in a truly random sequence.

The alternative to randomness is often simply nonrandomness. In a test based on the total number of runs, both too few and too many runs suggest lack of randomness. A null hypothesis of randomness would consequently be rejected if the total number of runs is either too large or too small. However, the two situations may indicate different types of lack of randomness. In the movie theater example, a sequence with too many runs, tending toward the genders alternating, might suggest that the movie is popular with teenagers and young adults, whereas the other extreme arrangement may result if the movie is more popular with younger children.

Tests of randomness are an important addition to statistical theory, because the theoretical bases for almost all the classical techniques, as well as distribution-free procedures, begin with the assumption of a random sample. If this assumption is valid, every sequential order is of no consequence. However, if the randomness of the observations is suspect, the information about order, which is almost always available, can be used to test a hypothesis of randomness. This kind of analysis is also useful in time-series and quality-control studies.

The symbols studied for pattern may arise naturally, as with the theater example, or may be artificially imposed according to some dichotomizing criterion. Thus the runs tests are applicable to either qualitative or quantitative data. In the latter case, the dichotomy is usually effected by comparing the magnitude of each number with a focal point, commonly the median or mean of the sample, and noting whether each observation exceeds or is exceeded by this value. When the data consist of numerical observations, two other types of runs analysis can be used to reach a conclusion about randomness. Both of these techniques use the information about relative magnitudes of adjacent numbers in the time-ordered sequence. These techniques, called the *runs up and down test* and the *rank von Neumann test*, use more of the available information and are especially effective when the alternative to randomness is either a trend or autocorrelation.

3.2 TESTS BASED ON THE TOTAL NUMBER OF RUNS

Assume an ordered sequence of n elements of two types, n_1 of the first type and n_2 of the second type, where $n_1 + n_2 = n$. If r_1 is the number of runs of type 1 elements and r_2 is the number of runs of type 2, the total number of runs in the sequence is $r = r_1 + r_2$. In order to derive a test for randomness based on the variable R, the total number of runs, we need the probability distribution of R when the null hypothesis of randomness is true.

EXACT NULL DISTRIBUTION OF *R*

The distribution of R will be found by first determining the joint probability distribution of R_1 and R_2 and then the distribution of their sum. Since under the null hypothesis every arrangement of the $n_1 + n_2$ objects is equiprobable, the probability that $R_1 = r_1$ and $R_2 = r_2$ is the number of distinguishable arrangements of $n_1 + n_2$ objects with r_1 runs of type 1 and r_2 runs of type 2 objects divided by the total number of distinguishable arrangements, which is $n!/n_1!n_2!$. For the numerator quantity, the following counting lemma can be used.

Lemma 1 *The number of distinguishable ways of distributing n-like objects into r distinguishable cells with no cell empty is* $\binom{n-1}{r-1}, n \geqslant r$.

Proof Suppose that the n-like objects are all white balls. Place these n balls in a row and effect the division into r cells by inserting each of $r - 1$ black balls between any two white balls in the line. Since there are $n - 1$ positions in which each black ball can be placed, the total number of arrangements is $\binom{n-1}{r-1}$.

In order to obtain a sequence with r_1 runs of the objects of type 1, the n_1-like objects must be placed into r_1 cells, which can be done in $\binom{n_1-1}{r_1-1}$ different ways. The same reasoning applies to obtain r_2 runs of the other n_2 objects. The total number of distinguishable arrangements starting with a run of type 1 then is the product $\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}$. Similarly for a sequence starting with a run of type 2. The blocks of objects of type 1 and type 2 must alternate, and consequently either $r_1 = r_2 \pm 1$ or $r_1 = r_2$. If $r_1 = r_2 + 1$, the sequence must begin with a run of type 1; if $r_1 = r_2 - 1$, a type 2 run must come first. But if $r_1 = r_2$, the sequence can begin with a run of either type, so the number of distinguishable arrangements must be doubled. We have thus proved the following result.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Theorem 2.1 *Let R_1 and R_2 denote the respective numbers of runs of n_1 objects of type 1 and n_2 objects of type 2 in a random sample of size $n = n_1 + n_2$. The joint probability distribution of R_1 and R_2 is*

$$f_{R_1,R_2}(r_1, r_2) = \frac{c\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n_1}} \qquad \begin{array}{l} r_1 = 1, 2, \ldots, n_1 \\ r_2 = 1, 2, \ldots, n_2 \\ r_1 = r_2 \text{ or } r_1 = r_2 \pm 1 \end{array} \tag{2.1}$$

where $c = 2$ if $r_1 = r_2$ and $c = 1$ if $r_1 = r_2 \pm 1$.

Corollary 2.1 *The marginal probability distribution of R_1 is*

$$f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1}}{\binom{n_1+n_2}{n_1}} \qquad r_1 = 1, 2, \ldots, n_1 \tag{2.2}$$

Similarly for R_2 with n_1 and n_2 interchanged.

Proof From (2.1), the only possible values of r_2 are $r_2 = r_1, r_1 - 1$, and $r_1 + 1$, for any r_1. Therefore we have

$$f_{R_1}(r_1) = \sum_{r_2} f_{R_1,R_2}(r_1, r_2)$$

$$\begin{aligned}\binom{n_1+n_2}{n_1} f_{R_1}(r_1) &= 2\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_1-1} + \binom{n_1-1}{r_1-1}\binom{n_2-1}{r_1-2} \\ &\quad + \binom{n_1-1}{r_1-1}\binom{n_2-1}{r_1} \\ &= \binom{n_1-1}{r_1-1}\left[\binom{n_2-1}{r_1-1} + \binom{n_2-1}{r_1-2} + \binom{n_2-1}{r_1-1}\right. \\ &\quad \left. + \binom{n_2-1}{r_1}\right] = \binom{n_1-1}{r_1-1}\left[\binom{n_2}{r_1-1} + \binom{n_2}{r_1}\right] \\ &= \binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1}\end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Theorem 2.2 *The probability distribution of R, the total number of runs of $n = n_1 + n_2$ objects, n_1 of type 1 and n_2 of type 2, in a random sample is*

$$f_R(r) = \begin{cases} 2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}\Big/\binom{n_1+n_2}{n_1} & \text{if } r \text{ is even} \\ \binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}+\binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}\Big/\binom{n_1+n_2}{n_1} & \text{if } r \text{ is odd} \end{cases} \quad (2.3)$$

for $r = 2, 3, \ldots, n_1 + n_2$.

Proof For r even, there must be the same number of runs of both types. Thus the only possible values of r_1 and r_2 are $r_1 = r_2 = r/2$, and (2.1) is summed over this pair. If $r_1 = r_2 \pm 1, r$ is odd. In this case, (2.1) is summed over the two pairs of values $r_1 = (r-1)/2$ and $r_2 = (r+1)/2, r_1 = (r+1)/2$ and $r_2 = (r-1)/2$, obtaining the given result. Note that the binomial coefficient $\binom{a}{b}$ is defined to be zero if $a < b$.

Using the result of Theorem 2.2, tables can be prepared for tests of significance of the null hypothesis of randomness. For example, if $n_1 = 5$ and $n_2 = 4$, we have

$$f_R(9) = \frac{\binom{4}{4}\binom{3}{3}}{\binom{9}{4}} = \frac{1}{126} = 0.008$$

$$f_R(8) = \frac{2\binom{4}{3}\binom{3}{3}}{\binom{9}{4}} = \frac{8}{126} = 0.063$$

$$f_R(2) = \frac{2\binom{4}{0}\binom{3}{0}}{\binom{9}{4}} = \frac{2}{126} = 0.016$$

$$f_R(3) = \frac{\binom{4}{1}\binom{3}{0}+\binom{4}{0}\binom{3}{1}}{\binom{9}{4}} = \frac{7}{126} = 0.056$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For a two-sided test which rejects the null hypothesis for $R \leqslant 2$ or $R \geqslant 9$, the exact significance level α would be $3/126 = 0.024$. For the critical region defined by $R \leqslant 3$ or $R \geqslant 8$, $\alpha = 18/126 = 0.143$. As for all tests based on discrete probability distributions, there are a finite number of possible values of α. For a test with significance level at most 0.05, say, the first critical region above would be used even though the actual value of α is only 0.024. Tables which can be used to find rejection regions for the runs test are available in many sources. Swed and Eisenhart (1943) give the probability distribution of R for $n_1 \leqslant n_2 \leqslant 20$, which is given in Table D of the Appendix for all $n_1 \leqslant n_2$ such that $n_1 + n_2 \leqslant 20$ and other n_1 and n_2 such that $n_1 \leqslant n_2 \leqslant 12$; note that left-tail and right-tail probabilities are given separately.

MOMENTS OF THE NULL DISTRIBUTION OF *R*

The kth moment of R is

$$\begin{aligned} E(R^k) &= \sum_r r^k f_R(r) \\ &= \Bigg\{ \sum_{r \text{ even}} 2r^k \binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1} \\ &\quad + \sum_{r \text{ odd}} r^k \Bigg[\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2} \\ &\quad + \binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2} \Bigg] \Bigg\} \Big/ \binom{n_1+n_2}{n_1} \end{aligned} \tag{2.4}$$

The smallest value of r is always 2. If $n_1 = n_2$, the largest number of runs occurs when the symbols alternate, in which case $r = 2n_1$. If $n_1 < n_2$, the maximum value of r is $2n_1 + 1$, since the sequence can both begin and end with a type 2 symbol. Assuming without loss of generality that $n_1 \leqslant n_2$, the range of summation for r is $2 \leqslant r \leqslant 2n_1 + 1$. Letting $r = 2i$ for r even and $r = 2i + 1$ for r odd, the range of i is $1 \leqslant i \leqslant n_1$.

For example, the mean of R is expressed as follows using (2.4):

$$\begin{aligned} &\binom{n_1+n_2}{n_1} E(R) \\ &= \sum_{i=1}^{n_1} 4i \binom{n_1-1}{i-1}\binom{n_2-1}{i-1} + \sum_{i=1}^{n_1} (2i+1)\binom{n_1-1}{i}\binom{n_2-1}{i-1} \\ &\quad + \sum_{i=1}^{n_1} (2i+1)\binom{n_1-1}{i-1}\binom{n_2-1}{i} \end{aligned} \tag{2.5}$$

To evaluate these three sums, the following lemmas are useful.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Lemma 2

$$\sum_{r=0}^{c}\binom{m}{r}\binom{n}{r}=\binom{m+n}{m} \qquad where\ c=\min(m,n)$$

Proof $(1+x)^{m+n}=(1+x)^m(1+x)^n$ for all x.

$$\sum_{i=0}^{m+n}\binom{m+n}{i}x^i=\sum_{j=0}^{m}\binom{m}{j}x^j\sum_{k=0}^{n}\binom{n}{k}x^k$$

Assuming without loss of generality that $c=m$ and equating the two coefficients of x^m on both sides of this equation, we obtain

$$\binom{m+n}{m}=\sum_{r=0}^{m}\binom{m}{m-r}\binom{n}{r}$$

Lemma 3

$$\sum_{r=0}^{c}\binom{m}{r}\binom{n}{r+1}=\binom{m+n}{m+1} \qquad where\ c=\min(m,n-1)$$

Proof The proof follows as in Lemma 2, equating coefficients of x^{m+1}.

The algebraic process of obtaining $E(R)$ from (2.5) is tedious and will be left as an exercise for the reader. The variance of R can be found by evaluating the factorial moment $E[R(R-1)]$ in a similar manner.

A much simpler approach to finding the moments of R is provided by considering R as the sum of indicator variables as follows for $n=n_1+n_2$. Let

$$R=1+I_2+I_3+\cdots+I_n$$

where in an ordered sequences of the two types of symbols, we define

$$I_k=\begin{cases}1 & \text{if the } k\text{th element} \neq \text{the } (k-1)\text{th element}\\ 0 & \text{otherwise}\end{cases}$$

Then I_k is a Bernoulli random variable with parameter $p=n_1n_2/\binom{n}{2}$, so

$$E(I_k)=E(I_k^2)=\frac{2n_1n_2}{n(n-1)}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Since R is a linear combination of these I_k, we have

$$E(R) = 1 + \sum_{k=2}^{n} E(I_k) = 1 + \frac{2n_1n_2}{n_1+n_2} \tag{2.6}$$

$$\text{var}(R) = \text{var}\left(\sum_{k=2}^{n} I_k\right) = (n-1)\text{var}(I_k) + \sum_{2 \leqslant j \neq k \leqslant n}\sum \text{cov}(I_j, I_k)$$

$$= (n-1)E(I_k^2) + \sum_{2 \leqslant j \neq k \leqslant n}\sum E(I_jI_k) - (n-1)^2[E(I_k)]^2 \tag{2.7}$$

To evaluate the $(n-1)(n-2)$ joint moments of the type $E(I_jI_k)$ for $j \neq k$, the subscript choices can be classified as follows:

1. For the $2(n-2)$ selections where $j = k-1$ or $j = k+1$,

$$E(I_jI_k) = \frac{n_1n_2(n_1-1) + n_2n_1(n_2-1)}{n(n-1)(n-2)} = \frac{n_1n_2}{n(n-1)}$$

2. For the remaining $(n-1)(n-2) - 2(n-2) = (n-2)(n-3)$ selections of $j \neq k$,

$$E(I_jI_k) = \frac{4n_1n_2(n_1-1)(n_2-1)}{n(n-1)(n-2)(n-3)}$$

Substitution of these moments in the appropriate parts of (2.7) gives

$$\text{var}(R) = \frac{2n_1n_2}{n} + \frac{2(n-2)n_1n_n}{n(n-1)} + \frac{4n_1n_2(n_1-1)(n_2-1)}{n(n-1)} - \frac{4n_1^2n_2^2}{n^2}$$

$$= \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \tag{2.8}$$

ASYMPTOTIC NULL DISTRIBUTION

Although (2.3) can be used to find the exact distribution of R for any values of n_1 and n_2, the calculations are laborious unless n_1 and n_2 are both small. For large samples an approximation to the null distribution can be used which gives reasonably good results as long as n_1 and n_2 are both larger than 10.

In order to find the asymptotic distribution, we assume that the total sample size n tends to infinity in such a way that $n_1/n \to \lambda$ and $n_2/n \to 1-\lambda, \lambda$ fixed, $0 < \lambda < 1$. For large samples then, the mean and variance of R from (2.6) and (2.8) are

$$\lim_{n\to\infty} E(R/n) = 2\lambda(1-\lambda) \qquad \lim_{n\to\infty} \operatorname{var}\left(R/\sqrt{n}\right) = 4\lambda^2(1-\lambda)^2$$

Forming the standardized random variable

$$Z = \frac{R - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)} \tag{2.9}$$

and substituting for R in terms of Z in (2.3), we obtain the standardized probability distribution of R, or $f_Z(z)$. If the factorials in the resulting expression are evaluated by Stirling's formula, the limit (Wald and Wolfowitz, 1940) is

$$\lim_{n\to\infty} \ln f_Z(z) = -\ln\sqrt{2\pi} - 1/2\,z^2$$

which shows that the limiting probability function of Z is the standard normal density.

For a two-sided test of size α using the normal approximation, the null hypothesis of randomness would be rejected when

$$\left|\frac{R - 2n\lambda(1-\lambda)}{2n^{1/2}\lambda(1-\lambda)}\right| \geqslant z_{\alpha/2} \tag{2.10}$$

where z_γ is that number which satisfies $\Phi(z_\gamma) = 1 - \gamma$ or, equivalently, z_γ is the $(1-\gamma)$th quantile (or the upper γth quantile) point of the standard normal probability distribution. The exact mean and variance of R given in (2.6) and (2.8) can also be used in forming the standardized random variable, as the asymptotic distribution is unchanged. These approximations are generally improved by using a continuity correction of 0.5, as explained in Chapter 1.

DISCUSSION

This runs test is one of the best known and easiest to apply among the tests for randomness in a sequence of observations. The data may be dichotomous as collected, or if actual measurements are collected the data may be classified into a dichotomous sequence according as each observation is above or below some fixed number, often the calculated sample median or mean. In this latter case, any observations equal to this fixed number are ignored in the analysis and n_1, n_2, and n reduced accordingly. The runs test can be used with either one- or two-sided alternatives. If the alternative hypothesis is simply randomness, a two-sided test should be used. Since the presence of a trend would usually be indicated by a clustering of like objects, which is reflected by an unusually small number of runs, a one-sided test is more appropriate for trend alternatives.

Because of the generality of alternatives to randomness, no statement can be made concerning the overall performance of this runs test. However, its versatility should not be underrated. Other tests for randomness have been proposed which are especially useful for trend alternatives. The best known of these are tests based on the length of the longest run and tests based on the runs up and down theory. These two types of test criteria will be discussed in the following sections.

APPLICATIONS

This test of randomness is based on the total number of runs R in an ordered sequence of n elements of two types, n_1 of type 1 and n_2 of type 2. Values of the null distribution of R computed from (2.3) are given in Table D for $n_1 \leqslant n_2 \leqslant 12$ as left-tail probabilities for R small and right-tail for R large. If the alternative is simply nonrandomness and the desired level is α, we should reject for $R \leqslant r_{\alpha/2}$ or $R \geqslant r'_{\alpha/2}$, where $P(R \leqslant r_{\alpha/2}) \leqslant \alpha/2$ and $P(R \geqslant r'_{\alpha/2}) \leqslant \alpha/2$; the exact level of this two-tailed test is $P(R \leqslant r_{\alpha/2}) + P(R \geqslant r'_{\alpha/2})$. If the desired alternative is a tendency for like elements to cluster, we should reject only for two few runs and therefore only for $R \leqslant r_{\alpha/2}$. On the other hand, if the appropriate alternative is a tendency for the elements to mix, we should reject only for too many runs and therefore only for $R \geqslant r'_{\alpha}$. Because Table D covers only $n_1 \leqslant n_2$, the type of element which occurs less frequently in the n observations should be called the type 1 element.

For $n_1 > 12$ and $n_2 > 12$, the critical values r and r' can be found from the normal approximation to the null distribution of the total number of runs. If we use the exact mean and variance of R given in (2.6) and (2.8) and a continuity correction of 0.5, the left-tail and right-tail critical regions are

$$\frac{R + 0.5 - 1 - 2n_1n_2/n}{\sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n-1)]}} \leqslant -z_\alpha$$
$$\frac{R - 0.5 - 1 - 2n_1n_2/n}{\sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n-1)]}} \geqslant z_\alpha \qquad (2.11)$$

The two-tailed critical region is a combination of the above with z_α replaced by $z_{\alpha/2}$.

Example 2.1 The recorded high temperature in a Florida resort town for each of 10 consecutive days during the month of January of this year is compared with the historical average high for the same days in previous years and noted as either above historical average (A) or below average (B). For the data A A B A B B A A A B, test the null hypothesis of random direction of deviation from average high temperature against the alternative of nonrandomness, using level 0.05.

Solution Since the data consist of six A's and four B's, B will be called the type 1 element to make $n_1 = 4, n_2 = 6$. The total number of runs observed is $R = 6$. Table D shows that $P(R \leqslant 2) = 0.010$ and $P(R \geqslant 9) = 0.024$, and these are the largest respective probabilities that do not exceed 0.025; the rejection region is then $R \leqslant 2$ or $R \geqslant 9$ with exact level 0.034. Our $R = 6$ does not fall into this region, so we do not reject the null hypothesis of randomness at the 0.05 level.

The STATXACT solution to Example 2.1 is shown below. The reader can verify using Table D that the exact right-tailed P value is $P(R \geqslant 6) = 0.595$ and the asymptotic P value from (2.11) with a continuity correction is $P(Z \geqslant -0.2107) = 0.5952$.

```
                **********************************************
                   STATXACT Solution to Example 2.1
                **********************************************

ONE SAMPLE RUNS TEST

Summary of Exact distribution of ONE SAMPLE RUNS TEST statistic

       Value of cutpoint =        10.00  ( User Defined )

Min          Max          Mean       Std-dev     Observed    Standardized
2.000       9.000        5.800        1.424        6.000         -0.2107

Asymptotic p-value :
       Pr { Test Statistic .GE.        6.000 }          =       0.5835
       Two-sided: 2 * One-sided                          =       1.0000

Exact p-value:
       Pr{ Test Statistic .GE.        6.000 }            =       0.5952
       Pr{ Test Statistic .EQ.        6.000 }            =       0.2857
       Pr{|Test Statistic-Mean| .GE. |Observed-Mean|} =       1.0000
```

3.3 TESTS BASED ON THE LENGTH OF THE LONGEST RUN

A test based on the total number of runs in an ordered sequence of n_1 objects of type 1 and n_2 objects of type 2 is only one way of using information about runs to detect patterns in the arrangement. Other statistics of interest are provided by considering the lengths of these runs. Since a run which is unusually long reflects a tendency for like objects to cluster and therefore possibly a trend, Mosteller (1941) has suggested a test for randomness based on the length of the longest run. Exact and asymptotic probability distributions of the numbers of runs of given length are discussed in Mood (1940).

The joint probability distribution of R_1 and R_2 derived in the previous section, for the numbers of runs of the two types of objects, disregards the individual lengths and is therefore of no use in this problem. We now need the joint probability distribution for a total of $n_1 + n_2$ random variables, representing the numbers of runs of all possible lengths for each of the two types of elements in the dichotomy. Let the random variables $R_{ij}, i = 1, 2; j = 1, 2, \ldots, n_i$, denote, respectively, the numbers of runs of objects of type i which are of length j. Then the following obvious relationships hold:

$$\sum_{j=1}^{n_i} j r_{ij} = n_i \qquad \text{and} \qquad \sum_{j=1}^{n_i} r_{ij} = r_i \qquad \text{for } i = 1, 2$$

The total number of arrangements of the $n_1 + n_2$ symbol is still $\binom{n_1+n_2}{n_1}$, and each is equally likely under the null hypothesis of randomness. We must compute the number of arrangements in which there are exactly r_{ij} runs of type i and length j, for all i and j. Assume that r_1 and r_2 are held fixed. The number of arrangements of the r_1 runs of type 1 which are composed of r_{1j} runs of length j for $j = 1, 2, \ldots, n_1$ is simply the number of permutations of the r_1 runs with r_{11} runs of length 1, r_{12} runs of length $2, \ldots, r_{1n_1}$ of length n_1, where within each category the runs cannot be distinguished. This number is $r_1!/\prod_{j=1}^{n_1} r_{1j}!$. The number of arrangements for the r_2 runs of type 2 objects is similarly $r_2!/\prod_{j=1}^{n_2} r_{2j}!$. If $r_1 = r_2 \pm 1$, the total number of permutations of the runs of both types of objects is the product of these two expressions, but if $r_1 = r_2$, since the sequence can begin with either type of object, this number must be doubled. Therefore the following theorem is proved.

Theorem 3.1 *Under the null hypothesis of randomness, the probability that the r_1 runs of n_1 objects of type 1 and r_2 runs of n_2 objects of type 2 consist of exactly r_{1j}, $j = 1, 2, \ldots, n_1$ and r_{2j}, $j = 1, 2, \ldots, n_2$ runs of length j, respectively, is*

$$f(r_{11}, \ldots, r_{1n_1}, r_{21}, \ldots, r_{2n_2}) = \frac{cr_1!r_2!}{\prod_{i=1}^{2}\prod_{j=1}^{n_i} r_{ij}! \binom{n_1+n_2}{n_1}} \tag{3.1}$$

where $c = 2$ if $r_1 = r_2$ and $c = 1$ if $r_1 = r_2 \pm 1$.

Combining the reasoning for Theorem 2.1 with that of Theorem 3.1, the joint distribution of the $n_1 + 1$ random variables R_2 and R_{1j}, $j = 1, 2, \ldots, n_1$, can be obtained as follows:

$$f(r_{11}, \ldots, r_{1n_1}, r_2) = \frac{cr_1!\binom{n_2-1}{r_2-1}}{\prod_{j=1}^{n_1} r_{1j}! \binom{n_1+n_2}{n_1}} \tag{3.2}$$

The result is useful when only the total number, not the lengths, of runs of type 2 objects is of interest. This joint distribution, when summed over the values for r_2, gives the marginal probability distribution of the lengths of the r_1 runs of objects of type 1.

Theorem 3.2 *The probability that the r_1 runs of n_1 objects of type 1 consist of exactly $r_{1j}, j = 1, 2, \ldots, n_1$, runs of length j, respectively, is*

$$f(r_{11}, \ldots, r_{1n_1}) = \frac{r_1!\binom{n_2+1}{r_1}}{\prod_{j=1}^{n_1} r_{1j}! \binom{n_1+n_2}{n_1}} \tag{3.3}$$

The proof is exactly the same as for the corollary to Theorem 2.1. The distribution of lengths of runs of type 2 objects is similar.

In the probability distributions given in Theorems 3.1 and 3.2, both r_1 and r_2, and r_1, respectively, were assumed to be fixed numbers. In other words, these are conditional probability distributions given the fixed values. If these are not to be considered as fixed, the conditional distributions are simply summed over the possible fixed values, since these are mutually exclusive.

Theorem 3.2 can be used to find the null probability distribution of a test for randomness based on K, the length of the longest run of type 1. For example, the probability that the longest in any number of runs of type 1 is of length k is

$$\sum_{r_1} \sum_{r_{11},\dots,r_{1k}} \sum \frac{r_1! \binom{n_2+1}{r_1}}{\prod_{j=1}^{k} r_{1j}! \binom{n_1+n_2}{n_1}} \tag{3.4}$$

where the sums are extended over all sets of nonnegative integers satisfying $\sum_{j=1}^{k} r_{1j} = r_1, \sum_{j=1}^{k} jr_{1j} = n_1, r_{1k} \geqslant 1, r_1 \leqslant n_1 - k + 1$, and $r_1 \leqslant n_2 + 1$. For example, if $n_1 = 5, n_2 = 6$, the longest possible run is of length 5. There can be no other runs, so that $r_1 = 1$ and $r_{11} = r_{12} = r_{13} = r_{14} = 0$, and

$$P(K=5) = \frac{\binom{7}{1}}{\binom{11}{5}} = \frac{7}{462}$$

Similarly, we can obtain

$$P(K=4) = \frac{2!}{1!1!} \frac{\binom{7}{2}}{\binom{11}{5}} = \frac{42}{462}$$

$$P(K=3) = \frac{\frac{3!}{2!1!}\binom{7}{3} + \frac{2!}{1!1!}\binom{7}{2}}{\binom{11}{5}} = \frac{147}{462}$$

$$P(K=2) = \frac{\frac{4!}{3!1!}\binom{7}{4} + \frac{3!}{1!2!}\binom{7}{3}}{\binom{11}{5}} = \frac{245}{462}$$

$$P(K=1) = \frac{5!}{5!} \frac{\binom{7}{5}}{\binom{11}{5}} = \frac{21}{462}$$

For a significance level of at most 0.05 when $n_1 = 5, n_2 = 6$, the null hypothesis of randomness is rejected when there is a run of type 1 elements of length 5. In general, the critical region would be the arrangements with at least one run of length t or more.

Theorem 3.1 must be used if the test is to be based on the length of the longest run of either type of element in the dichotomy. These two theorems are tedious to apply unless n_1 and n_2 are both quite small. Tables are available in Bateman (1948) and Mosteller (1941).

Tests based on the length of the longest run may or may not be more powerful than a test based on the total number of runs, depending on the basis for comparison. Both tests use only a portion of the information available, since the total number of runs, although affected by the lengths of the runs, does not directly make use of information regarding these lengths, and the length of the longest run only partially reflects both the lengths of other runs and the total number of runs. Power functions are discussed in Bateman (1948) and David (1947).

3.4 RUNS UP AND DOWN

When numerical observations are available and the sequence of numbers is analyzed for randomness according to the number or lengths of runs of elements above and below the median, some information is lost which might be useful in identifying a pattern in the time-ordered observations. With runs up and down, instead of using a single focal point for the entire series, the magnitude of each element is compared with that of the immediately preceding element in the time sequence. If the next element is larger, a run up is started; if smaller, a run down is started. We can observe when the sequence increases, and for how long, when it decreases, and for how long. A decision concerning randomness then might be based on the number and lengths of these runs, whether up or down, since a large number of long runs should not occur in a truly random set of numerical observations. Since an excessive number of long runs is usually indicative of some sort of trend exhibited by the sequence, this type of analysis should be most sensitive to trend alternatives.

If the time-ordered observations were 8, 13, 1, 3, 4, 7, there is a run up of length 1, followed by a run down of length 1, followed by a run up of length 3. The sequence of six observations can be represented by five plus and minus signs, $+, -, +, +, +$, indicating their relative magnitudes. More generally, suppose there are n numbers,

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

no two alike, say $a_1 < a_2 < \cdots < a_n$ when arranged in order of magnitude. The time-ordered sequence of observations $S_n = (x_1, x_2, \ldots, x_n)$ represents some permutation of these n numbers. These are $n!$ permutations, each one representing a possible set of sample observations. Under the null hypothesis of randomness, each of these $n!$ arrangements is equally likely to occur. The test for randomness using runs up and down for the sequence S_n of dimension n is based on the derived sequence D_{n-1} of dimension $n-1$, whose ith element is the sign of the difference $x_{i+1} - x_i$, for $i = 1, 2, \ldots, n-1$. Let R_i denote the number of runs, either up or down, of length exactly i in the sequence D_{n-1} or S_n. We have the obvious restrictions $1 \leqslant i \leqslant n-1$ and $\sum_{i=1}^{n-1} ir_i = n-1$. The test for randomness will reject the null hypothesis when there are at least r runs of length t or more, where r and t are determined by the desired significance level. Therefore we must find the joint distribution of $R_1, R_2, \ldots, R_{n-1}$ under the null hypothesis when every arrangement S_n is equally likely. Let $f_n(r_{n-1}, r_{n-2}, \ldots, r_1)$ denote the probability that there are exactly r_{n-1} runs of length $n-1, \ldots, r_i$ runs of length $i, \ldots, r_1$ runs of length 1. If $u_n(r_{n-1}, \ldots, r_1)$ represents the corresponding frequency, then

$$f_n = \frac{u_n}{n!}$$

because there are $n!$ possible arrangements of S_n. Since the probability distribution will be derived as a recursive relation, let us first consider the particular case where $n = 3$ and see how the distribution for $n = 4$ can be generated from it.

Given three numbers $a_1 < a_2 < a_3$, only runs of lengths 1 and 2 are possible. The $3! = 6$ arrangements and their corresponding values of r_2 and r_1 are given in Table 4.1. Since the probability of at least one run of length 2 or more is 2/6, if this defines the critical region the significance level is 0.33. With this size sample, a smaller significance

Table 4.1 Listing of arrangements when $n = 3$

S_3	D_2	r_2	r_1	*Probability distribution*
(a_1, a_2, a_3)	$(+, +)$	1	0	
(a_1, a_3, a_2)	$(+, -)$	0	2	$f_3(1, 0) = 2/6$
(a_2, a_1, a_3)	$(-, +)$	0	2	
(a_2, a_3, a_1)	$(+, -)$	0	2	$f_3(0, 2) = 4/6$
(a_3, a_1, a_2)	$(-, +)$	0	2	
(a_3, a_2, a_1)	$(-, -)$	1	0	$f_3(r_2, r_1) = 0$ otherwise

level cannot be obtained without resorting to a randomized decision rule.

Now consider the addition of a fourth number a_4, larger than all the others. For each of the arrangements in S_3, a_4 can be inserted in four different places. In the particular arrangement (a_1, a_2, a_3), for example, insertion of a_4 at the extreme left or between a_2 and a_3 would leave r_2 unchanged but add a run of length 1. If a_4 is placed at the extreme right, a run of length 2 is increased to a run of length 3. If a_4 is inserted between a_1 and a_2, the one run of length 2 is split into three runs, each of length 1.

Extending this analysis to the general case, the extra observation must either split an existing run, lengthen an existing run, or introduce a new run of length 1. The ways in which the run lengths in S_{n-1} are affected by the insertion of an additional observation a_n to make an arrangement S_n can be classified into the following four mutually exclusive and exhaustive cases:

1. An additional run of length 1 can be added in the arrangement S_n.
2. A run of length $i-1$ in S_{n-1} can be changed into a run of length i in S_n for $i = 2, 3, \ldots, n-1$.
3. A run of length $h = 2i$ in S_{n-1} can be split into a run of length i, followed by a run of length 1, followed by a run of length i, for $1 \leqslant i \leqslant [(n-2)/2]$, where $[x]$ denotes the largest integer not exceeding x.
4. A run of length $h = i + j$ in S_{n-1} can be split up into
   a. A run of length i, followed by a run of length 1, followed by a run of length j
   b. A run of length j, followed by a run of length 1, followed by a run of length i.

where $h > i > j, 3 \leqslant h \leqslant n-2$.

For $n = 4$, the arrangements can be enumerated systematically in a table like Table 4.2 to show how these cases arise. Table 4.2 gives a partial listing. When the table is completed the number of cases which result in any particular set (r_3, r_2, r_1) can be counted and divided by 24 to obtain the complete probability distribution. This will be left as an exercise for the reader. The results are:

(r_3, r_2, r_1)	(1,0,0)	(0,1,1)	(0,0,3)	Other values
$f_4(r_3, r_2, r_1)$	2/24	12/24	10/24	0

Table 4.2 Partial listing of arrangements when $n = 4$

S_3	r_2	r_1	S_4	r_3	r_2	r_1	*Case illustrated*
(a_1, a_2, a_3)	1	0	(a_4, a_1, a_2, a_3)	0	1	1	1
			(a_1, a_4, a_2, a_3)	0	0	3	3
			(a_1, a_2, a_4, a_3)	0	1	1	1
			(a_1, a_2, a_3, a_4)	1	0	0	2
(a_1, a_3, a_2)	0	2	(a_4, a_1, a_3, a_2)	0	0	3	1
			(a_1, a_4, a_3, a_2)	0	1	1	2
			(a_1, a_3, a_4, a_2)	0	1	1	2
			(a_1, a_3, a_2, a_4)	0	0	3	1

There is no illustration for case 4 in the completed table of enumerated arrangements since here n is not large enough to permit $h \geqslant 3$. For $n = 5$, insertion of a_5 in the second position of (a_1, a_2, a_3, a_4) produces the sequence $(a_1, a_5, a_2, a_3, a_4)$. The one run of length 3 has been split into a run of length 1, followed by another run of length 1, followed by a run of length 2, illustrating case $4b$ with $h = 3,\ j = 1,\ i = 2$. Similarly, case $4a$ is illustrated by inserting a_5 in the third position. This also illustrates Case 3.

More generally, the frequency u_n of cases in S_n having exactly r_1 runs of length 1, r_2 runs of length $2, \ldots, r_{n-1}$ runs of length $n-1$ can be generated from the frequencies for S_{n-1} by the following recursive relation:

$$
\begin{aligned}
&u_n(r_{n-1}, r_{n-2}, \ldots, r_h, \ldots, r_i, \ldots, r_j, \ldots, r_1) \\
&\quad = 2u_{n-1}(r_{n-2}, \ldots, r_1 - 1) \\
&\qquad + \sum_{i=2}^{n-1} (r_{i-1} + 1) u_{n-1}(r_{n-2}, \ldots, r_i - 1, r_{i-1} + 1, \ldots, r_1) \\
&\qquad + \sum_{\substack{i=1 \\ (h=2i)}}^{[(n-2)/2]} (r_h + 1) u_{n-1}(r_{n-2}, \ldots, r_h + 1, \ldots, r_i - 2, \ldots, r_1 - 1) \\
&\qquad + 2 \sum_{i=2}^{n-3} \sum_{\substack{j=1 \\ (h=i+j) \\ h \leqslant n-2}}^{i-1} (r_h + 1) \\
&\qquad \times u_{n-1}(r_{n-2}, \ldots, r_h + 1, \ldots, r_i - 1, \ldots, r_j - 1, \ldots, r_1 - 1) \qquad (4.1)
\end{aligned}
$$

The terms in this sum represent cases 1 to 4 in that order. For case 1, u_{n-1} is multiplied by 2 since for every arrangement in S_{n-1}

there are exactly two places in which a_n can be inserted to add a run of length 1. These positions are always at an end or next to the end. If the first run is a run up (down), insertion of a_n at the extreme left (next to extreme left) position adds a run of length 1. A new run of length 1 is also created by inserting a_n at the extreme right (next to extreme right) in S_{n-1} if the last run in S_{n-1} was a run down (up). In case 4, we multiply by 2 because of the (a) and (b) possibilities.

The result is tricky and tedious to use but is much easier than enumeration. The process will be illustrated for $n = 5$, given

$$u_4(1,0,0) = 2 \qquad u_4(0,1,1) = 12 \qquad u_4(0,0,3) = 10$$
$$u_4(r_3, r_2, r_1) = 0 \quad \text{otherwise}$$

Using (4.1), we have

$$\begin{aligned}
u_5(r_4, r_3, r_2, r_1) &= 2u_4(r_3, r_2, r_1 - 1) + [(r_1 + 1)u_4(r_3, r_2 - 1, r_1 + 1)] \\
&\quad + (r_2 + 1)u_4(r_3 - 1, r_2 + 1, r_1) \\
&\quad + (r_3 + 1)u_4(r_3 + 1, r_2, r_1) \\
&\quad + (r_2 + 1)u_4(r_3, r_2 + 1, r_1 - 3) \\
&\quad + 2(r_3 + 1)u_4(r_3 + 1, r_2 - 1, r_1 - 2) \\
u_5(1,0,0,0) &= 1u_4(1,0,0) = 2 \\
u_5(0,1,0,1) &= 2u_4(1,0,0) + 1u_4(0,1,1) \\
&\quad + 2u_4(2,0,1) = 4 + 12 + 0 = 16 \\
u_5(0,0,2,0) &= 1u_4(0,1,1) + 1u_4(1,2,0) = 12 + 0 = 12 \\
u_5(0,0,1,2) &= 2u_4(0,1,1) + 3u_4(0,0,3) + 1u_4(1,1,2) \\
&\quad + 2u_4(1,0,0) = 24 + 30 + 0 + 4 = 58 \\
u_5(0,0,0,4) &= 2u_4(0,0,3) + 1u_4(1,0,4) + 1u_4(0,1,1) \\
&= 20 + 0 + 12 = 32
\end{aligned}$$

The means, variances, and covariances of the numbers of runs of length t (or more) are found in Levene and Wolfowitz (1944). Tables of the exact probabilities of at least r runs of length t or more are given in Olmstead (1946) and Owen (1962) for $n \leqslant 14$, from which appropriate critical regions can be found. Olmstead gives approximate probabilities for larger sample sizes. See Wolfowitz (1944a,b) regarding the asymptotic distribution which is Poisson.

A test for randomness can also be based on the total number of runs, whether up or down, irrespective of their lengths. Since the total number of runs R is related to the R_i, the number of runs of length i, by

$$R = \sum_{i=1}^{n-1} R_i \tag{4.2}$$

the same recursive relation given in (4.1) can be used to find the probability distribution of R. Levene (1952) showed that the asymptotic distribution of the standardized random variable R with mean $(2n-1)/3$ and variance $(16n-29)/90$ is the standard normal distribution.

APPLICATIONS

For the test of randomness based on the total number of runs up and down, R, in an ordered sequence of n numerical observations, or equivalently a sequence of $n-1$ plus and minus signs, the appropriate sign is determined by comparing the magnitude of each observation with the one immediately preceding it in the sequence. The appropriate rejection regions for each alternative are exactly the same as for the earlier test in Section 3.2, which was based on the total number of runs of two types of elements. Specifically, if the alternative is a tendency for like signs to cluster, the appropriate rejection region is small values of R. If the alternative is a tendency for like signs to mix with unlike signs, the appropriate rejection region is large values of R.

The exact distribution of R under the null hypothesis of randomness is given in Table E for $n \leqslant 25$ as left-tail probabilities for R small and right-tail for R large. For $n > 25$, the critical values of R can be found from the normal approximation to the null distribution of the number of runs up and down statistic. If we incorporate a continuity correction of 0.5, the left-tail and right-tail critical regions are

$$\frac{R+0.5-(2n-1)/3}{\sqrt{(16n-29)/90}} \leqslant -z_\alpha \quad \text{and} \quad \frac{R-0.5-(2n-1)/3}{\sqrt{(16n-29)/90}} \geqslant z_\alpha$$

The two-tailed critical region is a combination of the above with z_α replaced by $z_\alpha/2$.

One of the primary uses of the runs up and down test is for an analysis of time series data. The null hypothesis of randomness is then interpreted as meaning the data can be regarded as independent and identically distributed. The alternative of a tendency to cluster is interpreted as an upward trend if the signs are

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

predominantly plus or a downward trend if the signs are predominantly minus, and the alternative of a tendency to mix is interpreted as cyclical variations. The total number of runs test could also be used with time series data if the data are first converted to two types of symbols by comparing the magnitude of each one to some standard for that period or to a single focal point, like the median of the data. The test in this situation is frequently referred to as the runs above and below the median test. The test in the former case was actually illustrated by Example 2.1.

Example 4.1 illustrates an application of runs up and down in time series data.

Example 4.1 Tourism is regarded by all nations as big business because the industry brings needed foreign exchange and helps the balance of payments. The Travel Market Yearbook publishes extensive data on tourism. Analyze the annual data on total number of tourists to the United States for 1970–1982 to see if there is evidence of a trend, using the 0.01 level.

Year	*Number of tourists (millions)*
1970	12,362
1971	12,739
1972	13,057
1973	13,955
1974	14,123
1975	15,698
1976	17,523
1977	18,610
1978	19,842
1979	20,310
1980	22,500
1981	23,080
1982	21,916

Solution We use the runs up and down test of randomness for these $n = 13$ observations with the alternative of a trend. The sequence of 12 plus and minus signs is $+, +, +, +, +, +, +, +, +, +, +, -$ so that $R = 2$. The left-tail critical value from Table E is $R = 4$ at exact level 0.0026, the largest value that does not exceed 0.01. Since 2 is less than 4, we reject the null hypothesis and conclude there is a trend in number of tourists to the United States and it is positive because the signs are predominantly plus.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

3.5 A TEST BASED ON RANKS

Another way to test for randomness by comparing the magnitude of each element with that of the immediately preceding element in the time sequence is to compute the sum of the squares of the deviations of the pairs of successive elements. If the magnitudes of these elements are replaced by their respective ranks in the sequence before computing the sum of squares of successive deviations, we can obtain a nonparametric test.

Specifically, let the time-ordered sequence of observations be $S_n = (X_1, X_2, \ldots, X_n)$ as in Section 3.4. The test statistic

$$\text{NM} = \sum_{i=1}^{n-1} [\text{rank}(X_i) - \text{rank}(X_{i+1})]^2 \tag{5.1}$$

was proposed by Bartels (1982). A test based on a function of this statistic is the rank version of the ratio test for randomness developed by von Neumann using normal theory and is a linear transformation of the rank serial correlation coefficient introduced by Wald and Wolfowitz (1943).

It is easy to show that the test statistic NM ranges between $(n-1)$ and $(n-1)(n^2+n-3)/3$ if n is even and between $(n-1)$ and $[(n-1)(n^2+n-3)/3] - 1$ if n is odd. The exact null distribution of NM can be found by enumeration and is given in Bartels (1982) for $4 \leqslant n \leqslant 10$. For larger sample sizes, the test statistic

$$\text{RVN} = \frac{\sum_{i=1}^{n-1} [\text{rank}(X_i) - \text{rank}(X_{i+1})]^2}{\sum_{i=1}^{n} [\text{rank}(X_i) - (n+1)/2]^2} \tag{5.2}$$

is asymptotically normally distributed with mean 2 and variance $4(n-2)(5n^2-2n-9)/5n(n+1)(n-1)^2$, which is approximately equal to $20/(5n+7)$. If there are no ties, the denominator of RVN is equal to the constant $n(n^2-1)/12$.

Since a trend in either direction will be reflected by a small value of NM and therefore RVN, the appropriate rejection region to test randomness against a trend alternative is small values of NM or RVN. If the alternative is a tendency for the data to alternate small and large values, the appropriate rejection region is large values of the test statistic. Table S of the Appending gives exact tail probabilities for selected values of NM for $n \leqslant 10$ and approximate left-tail critical values of RVN (based on the beta distribution) for larger sample sizes. Corresponding right-tail critical values are

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

found using the fact that this beta approximation is symmetric about 2.0.

Bartels (1982) used simulation studies to show that this test is superior to the runs up and down test in many cases. Its asymptotic relative efficiency is 0.91 with respect to the ordinary serial correlation coefficient against the alternative of first-order autocorrelation under normality. Autocorrelation is defined as a measure of the dependence between observations that occur a fixed number of time units apart. Positive autocorrelation shows up in time series data that exhibit a trend in either direction, while negative autocorrelation is indicated by fluctuations over time.

Example 5.1 We illustrate this rank Von Neumann test using the data from Example 4.1 with $n = 13$ where the alternative is a positive trend. We first rank the number of tourists from smallest to largest and obtain

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 11$$

The value of NM, the numerator of the RVN statistic, from (5.1) is then

$$\text{NM} = (1-2)^2 + (2-3)^2 + \cdots + (13-11)^2 = 18$$

and the denominator is $13(13^2 - 1)/12 = 182$. Thus from (5.2), $\text{RVN} = 18/182 = 0.0989$, and Table S shows that the left-tail critical value based on the beta approximation is to reject the null hypothesis of randomness at the 0.005 level if $\text{RVN} \leqslant 0.74$. Therefore we reject the null hypothesis and conclude that there is a significant positive trend. We also use these data to illustrate the test based on the normal approximation to the distribution of RVN. The mean is 2 and the variance for $n = 13$ is 0.2778. The standard normal test statistic is then $(0.0989 - 2)/\sqrt{0.2778} = -3.61$. At the 0.005 level, for example, the appropriate rejection region from Table A is $Z \leqslant -2.58$, so we again conclude that there is a significant postive trend.

3.6 SUMMARY

In this chapter we presented a number of tests that are appropriate for testing the null hypothesis of randomness in a sequence of observations whose order has some meaning. If the observations are simply two types of symbols, like M and F, or D and G, the total number of runs of symbols is the most appropriate test statistic. Tests based on the lengths of the runs are primarily of theoretical interest. If the

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

observations are numerical measurements, the number of runs up and down or the rank von Neumann (RVN) statistic provides the best test because too much information is lost by using the test based on runs above and below some fixed value like the median.

Usually the alternative hypothesis is simply lack of randomness, and then these tests have no analog in parametric statistics. If the data represent a time series, the alternative to randomness can be exhibition of a trend. In this case the RVN test is more powerful than the test based on runs up and down. Two additional tests for trend will be presented later in Chapter 11.

PROBLEMS

3.1. Prove Corollary 2.1 using a direct combinatorial argument based on Lemma 1.

3.2. Find the mean and variance of the number of runs R_1 of type 1 elements, using the probability distribution given in (2.2). Since $E(R) = E(R_1) + E(R_2)$, use your result to verify (2.6).

3.3. Use Lemmas 2 and 3 to evaluate the sums in (2.5), obtaining the result given in (2.6) for $E(R)$.

3.4. Show that the asymptotic distribution of the standardized random variable $[R_1 - E(R_1)]/\sigma(R_1)$ is the standard normal distribution, using the distribution of R_1 given in (2.2) and your answer to Problem 3.2.

3.5. Verify that the asymptotic distribution of the random variable given in (2.9) is the standard normal distribution.

3.6. By considering the ratios $f_R(r)/f_R(r-2)$ and $f_R(r+2)/f_R(r)$, where r is an even positive integer and $f_R(r)$ is given in (2.3) show that if the most probable number of runs is an even integer k, then k satisfies the inequality

$$\frac{2n_1n_2}{n} < k < \frac{2n_1n_2}{n} + 2$$

3.7. Show that the probability that a sequence of n_1 elements of type 1 and n_2 elements of type 2 begins with a type 1 run of length exactly k is

$$\frac{(n_1)_k n_2}{(n_1+n_2)_{k+1}} \qquad \text{where } (n)_r = \frac{n!}{(n-r)!}$$

3.8. Find the rejection region with significance level not exceeding 0.10 for a test of randomness based on the length of the longest run when $n_1 = n_2 = 6$.

3.9. Find the complete probability distribution of the number of runs up and down of various length when $n = 6$ using (4.1) and the results given for $u_5(r_4, r_3, r_2, r_1)$.

3.10. Use your answers to Problem 3.9 to obtain the complete probability distribution of the total number of runs up and down when $n = 6$.

3.11. Verify the statement that the variance of the RVN test statistic is approximately equal to $20/(5n+7)$.

3.12. Analyze the data in Example 4.1 for evidence of trend using total number of runs above and below

(*a*) The sample median

(*b*) The sample mean

3.13. A certain broker noted the following number of bonds sold each month for a 12-month period:

Jan. 19	July 22
Feb. 23	Aug. 24
Mar. 20	Sept. 25
Apr. 17	Oct. 28
May 18	Nov. 30
June 20	Dec. 21

(*a*) Use the runs up and down test to see if these data show a directional trend and make an appropriate conclusion at the 0.05 level.

(*b*) Use the runs above and below the sample median test to see if these data show a trend and make an appropriate conclusion at the 0.05 level.

(*c*) Compare the conclusions reached in (*a*) and (*b*) and give an explanation for the difference.

3.14. The following are 30 time lapses in minutes between eruptions of Old Faithful geyser in Yellowstone National Park, recorded between the hours of 8 a.m. and 10 p.m. on a certain day, and measured from the beginning of one eruption to the beginning of the next:

68, 63, 66, 63, 61, 44, 60, 62, 71, 62, 62, 55, 62, 67, 73,

72, 55, 67, 68, 65, 60, 61, 71, 60, 68, 67, 72, 69, 65, 66

A researcher wants to use these data for inference purposes, but is concerned about whether it is reasonable to treat such data as a random sample. What do you think? Justify your answer.

3.15. In a psychological experiment, the research question of interest is whether a rat "learned" its way through a maze during 65 trials. Suppose the time-ordered observations on number of correct choices by the rat on each trail are as follows:

0, 1, 2, 1, 1, 2, 3, 2, 2, 2, 1, 1, 3, 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 4, 3, 1, 2, 2, 1, 2, 2,

2, 2, 3, 2, 2, 3, 4, 3, 2, 3, 3, 2, 3, 3, 2, 3, 3, 2, 3, 4, 3, 3, 4, 2, 3, 3, 4, 3, 4, 4, 4, 4

(*a*) Test these data for randomness against the alternative of a tendency to cluster, using the dichotomizing criterion that 0, 1, or 2 correct choices indicate no learning, while 3 or 4 correct choices indicate learning.

(*b*) Would the runs up and down test be appropriate for these data? Why or why not?

3.16. The data below represent departure of actual daily temperature in degrees Fahrenheit from the normal daily temperature at noon at a certain airport on seven consecutive days.

Day	1	2	3	4	5	6	7
Departure	12	13	12	11	5	−1	−2

(*a*) Give an appropriate *P* value that reflects whether the pattern of positive and negative departures can be considered a random process or exhibits a tendency to cluster.

(*b*) Given an appropriate *P* value that reflects whether the pattern of successive departures (from one day to the next) can be considered a random process or exhibits a trend for these seven days.

3.17. The three graphs in Figure 1 (see below) illustrate some kinds of nonrandom patterns. Time is on the horizontal axis. The data values are indicated by dots and the horizontal line denotes the median of the data. For each graph, compute the one-tailed *P*-value for non randomness using two different nonparametric techniques.

3.18. Bartels (1982) illustrated the rank non Neumann test for randomness using data on annual changes in stock levels of corporate trading enterprises in Australia for 1968–1969 to 1977–1978. The values (in \$A million) deflated by the Australian GDP are 528, 348, 264, −20, −167, 575, 410, −4, 430, −122. He tested randomness against the alternative of autocorrelation. Random stock level changes occur when

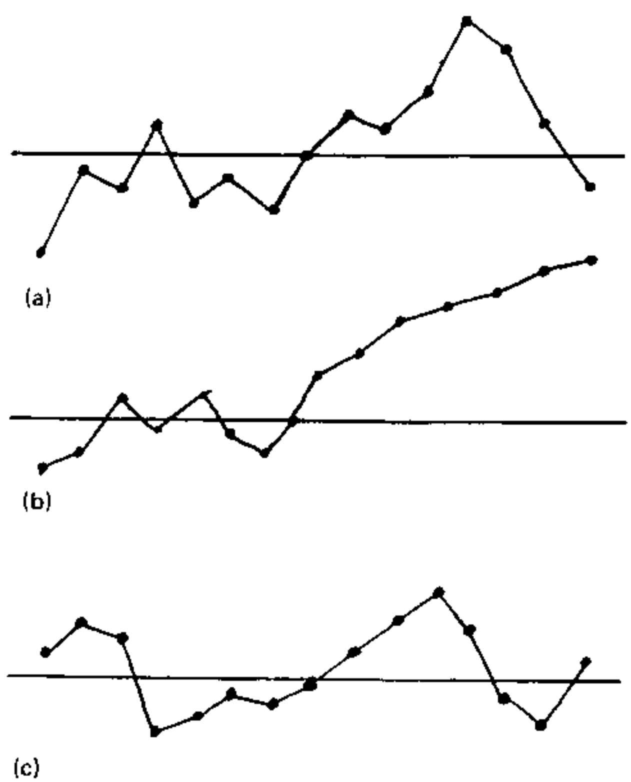


Fig. 1 Nonrandom patterns representing (a) cyclical movement, (b) trend movement, (c) clustering.

companies are well managed because future demands are accurately anticipated. "Negative autocorrelation constitutes evidence for a tendency to overreact to shortfalls or excesses in stock levels, whereas positive autocorrelation suggests there is a long delay in reaching desired stock levels." The test statistic is NM = 169, which is not significant. Compare this result with that of (*a*) runs up and down, and (*b*) with runs above and below the sample median.

4

Tests of Goodness of Fit

4.1 INTRODUCTION

An important problem in statistics relates to obtaining information about the form of the population from which a sample is drawn. The shape of this distribution might be the focus of the investigation. Alternatively, some inference concerning a particular aspect of the population may be of primary interest. In this latter case, in classical statistics, information about the form generally must be postulated or incorporated in the null hypothesis to perform an exact parametric type of inference. For example, suppose we have a small number of observations from an unknown population with unknown variance and the hypothesis of interest concerns the value of the population mean. The traditional parametric test, based on Student's t distribution, is derived under the assumption of a normal population. The exact distribution theory and probabilities of both types of errors depend on this population form. Therefore it might be desirable to check on the

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

reasonableness of the normality assumption before forming any conclusions based on the t distribution. If the normality assumption appears not to be justified, some type of nonparametric inference for location might be more appropriate with a small sample size.

The compatibility of a set of observed sample values with a normal distribution or any other distribution can be checked by a goodness-of-fit type of test. These are tests designed for a null hypothesis which is a statement about the form of the cumulative distribution function or probability function of the parent population from which the sample is drawn. Ideally, the hypothesized distribution is completely specified, including all parameters. Since the alternative is necessarily quite broad, including differences only in location, scale, other parameters, form, or any combination thereof, rejection of the null hypothesis does not provide much specific information. Goodness-of-fit tests are customarily used when only the form of the population is in question, with the hope that the null hypothesis will be found acceptable.

In this chapter we shall consider two types of goodness-of-fit tests. The first type is designed for null hypotheses concerning a discrete distribution and compares the observed frequencies with the frequencies expected under the null hypothesis. This is the chi-square test proposed by Karl Pearson early in the history of statistics. The second type of goodness-of-fit test is designed for null hypotheses concerning a continuous distribution and compares the observed cumulative relative frequencies with those expected under the null hypotheses. This group includes the Kolmogorov-Smirnov and the Lilliefors's tests. The latter is designed for testing the assumption of a normal or an exponential distribution with unspecified parameters and is therefore an important preliminary test for justifying the use of parametric or classical statistical methods that require this assumption. Finally, we present some graphical approaches to assessing the form of a distribution.

4.2 THE CHI-SQUARE GOODNESS-OF-FIT TEST

A single random sample of size n is drawn from a population with unknown cumulative distribution function F_X. We wish to test the null hypothesis

$$H_0: F_X(x) = F_0(x) \qquad \text{for all } x$$

where $F_0(x)$ is completely specified, against the general alternative

$$H_1: F_X(x) \neq F_0(x) \qquad \text{for some } x$$

In order to apply the chi-square test in this situation, the sample data must first be grouped according to some scheme in order to form a frequency distribution. In the case of count or qualitative data, where the hypothesized distribution would be discrete, the categories would be the relevant verbal or numerical classifications. For example, in tossing a die, the categories would be the numbers of spots; in tossing a coin, the categories would be the numbers of heads; in surveys of brand preferences, the categories would be the brand names considered. When the sample observations are quantitative, the categories would be numerical classes chosen by the experimenter. In this case, the frequency distribution is not unique and some information is necessarily lost. Even though the hypothesized distribution is most likely continuous with measurement data, the data must be categorized for analysis by the chi-square test.

When the population distribution is completely specified by the null hypothesis, one can calculate the probability that a random observation will be classified into each of the chosen or fixed categories. These probabilities multiplied by n give the frequencies for each category which would be expected if the null hypothesis were true. Except for sampling variation, there should be close agreement between these expected and observed frequencies if the sample data are compatible with the specified $F_0(x)$. The corresponding observed and expected frequencies can be compared visually using a histogram, a frequency polygon, or a bar chart. The chi-square goodness-of-fit test provides a probability basis for effecting the comparison and deciding whether the lack of agreement is too great to have occurred by chance.

Assume that the n observations have been grouped into k mutually exclusive categories, and denote the observed and expected frequencies for the ith class by f_i and e_i, respectively, $i = 1, 2, \ldots, k$. The decision regarding fit is to be based on the deviations $f_i - e_i$. The sum of these k deviations is zero except for rounding. The test criterion suggested by Pearson (1900) is the sum of squares of these deviations, normalized by the expected frequency, or

$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{2.1}$$

A large value of Q would reflect an incompatibility between the observed and expected relative frequencies, and therefore the null hypothesis used to calculate the e should be rejected for Q large.

The exact probability distribution of the random variable Q is quite complicated, but for large samples its distribution is approximately chi square with $k-1$ degrees of freedom, given here as Table B of the Appendix. The theoretical basis for this can be argued briefly as follows.

The only random variables of concern are the class frequencies $F_1, F_2, \ldots, F_k$, which constitute a set of random variables from the k-variate multinomial distribution with k possible outcomes, the ith outcome being the ith category in the classification system. With $\theta_1, \theta_2, \ldots, \theta_k$ denoting the probabilities of the respective outcomes and $f_1, f_2, \ldots, f_k$ denoting the observed outcomes, the likelihood function of the sample then is

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{k} \theta_i^{f_i} \qquad f_i = 0, 1, \ldots, n; \ \sum_{i=1}^{k} f_i = n; \ \sum_{i=1}^{k} \theta_i = 1 \tag{2.2}$$

The null hypothesis was assumed to specify the population distribution completely, from which the θ_i can be calculated. This hypothesis then is actually concerned only with the values of these parameters and can be equivalently stated as

$$H_0 \colon \theta_i^0 = \frac{e_i}{n} \qquad \text{for } i = 1, 2, \ldots, k$$

It is easily shown that the maximum-likelihood estimates of the parameters in (2.2) are $\hat{\theta}_i = f_i / n$. The likelihood-ratio statistic for this hypothesis then is

$$T = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{L(\theta_1^0, \theta_2^0, \ldots, \theta_k^0)}{L(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)} = \prod_{i=1}^{k} \left(\frac{\theta_i^0}{\hat{\theta}_i} \right)^{f_i}$$

As was stated in Chapter 1, the distribution of the random variable $-2 \ln T$ can be approximated by the chi-square distribution. The degrees of freedom are $k-1$, since the restriction $\sum_{i=1}^{k} \theta_i = 1$ leaves only $k-1$ parameters in Ω to be estimated independently. We have here

$$-2 \ln T = -2 \sum_{i=1}^{k} f_i \left(\ln \theta_i^0 - \ln \frac{f_i}{n} \right) \tag{2.3}$$

Some statisticians advocate using the expression in (2.3) as a test criterion for goodness-of-fit. We shall now show that it is asymptotically equivalent to the expression for Q given in (2.1). The Taylor series expansion of $\ln \theta_i$ about $f_i / n = \hat{\theta}_i$ is

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$\ln \theta_i = \ln \hat{\theta}_i + (\theta_i - \hat{\theta}_i)\frac{1}{\hat{\theta}_i} + \frac{(\theta_i - \hat{\theta}_i)^2}{2!}\left(-\frac{1}{\hat{\theta}_i^2}\right) + \epsilon$$

so that

$$\ln \theta_i^0 - \ln \frac{f_i}{n} = \left(\theta_i^0 - \frac{f_i}{n}\right)\frac{n}{f_i} - \left(\theta_i^0 - \frac{f_i}{n}\right)^2 \frac{n^2}{2f_i^2} + \epsilon$$

$$= \frac{(n\theta_i^0 - f_i)}{f_i} - \frac{(n\theta_i^0 - f_i)^2}{2f_i^2} + \epsilon$$

where ϵ represents the sum of terms alternating in sign

$$\sum_{j=3}^{\infty}(-1)^{j+1}\left(\theta_i^0 - \frac{f_i}{n}\right)^j \frac{n^j}{j!f_i^j} \tag{2.4}$$

Substituting (2.4) in (2.3), we have

$$-2\ln T = -2\sum_{i=1}^{k}(n\theta_i^0 - f_i) + \sum_{i=1}^{k}\frac{(n\theta_i^0 - f_i)^2}{f_i} + \sum_{i=1}^{k}\epsilon'$$

$$= 0 + \sum_{i=1}^{k}\frac{(f_i - e_i)^2}{f_i} + \epsilon''$$

By the law of large numbers F_i/n is known to be a consistent estimator of θ_i, or

$$\lim_{n\to\infty}\left[\frac{1}{n}P(|F_i - n\theta_i| > \epsilon)\right] = 0 \qquad \text{for every } \epsilon > 0$$

Thus we see that the probability distribution of Q converges to that of $-2\ln T$, which is chi square with $k-1$ degrees of freedom. An approximate α-level test then is obtained by rejecting H_0 when Q exceeds the $(1-\alpha)$th quantile point of the chi-square distribution, denoted by $\chi^2_{k-1,\alpha}$. This approximation can be used with confidence as long as every expected frequency is at least equal to 5. For any e_i smaller than 5, the usual procedure is to combine adjacent groups in the frequency distribution until this restriction is satisfied. The number of degrees of freedom then must be reduced to correspond to the actual number of categories used in the analysis. This rule of 5 should not be considered inflexible, however. It is conservative, and the chi-square approximation is often reasonably accurate for expected cell frequencies as small as 1.5.

Any case where the θ_i are completely specified by the null hypothesis is thus easily handled. The more typical situation, however, is where the null hypothesis is composite, i.e., it states the form of the distribution but not all the relevant parameters. For example, when we wish to test whether a sample is drawn from some normal population, μ and σ would not be given. However, in order to calculate the expected frequencies under H_0, μ and σ must be known. If the expected frequencies are estimated from the data as $n\hat{\theta}_i^0$ for $i = 1, 2, \ldots, k$, the random variable for the goodness-of-fit test in (2.1) becomes

$$Q = \sum_{i=1}^{k} \frac{(F_i - n\hat{\theta}_i^0)^2}{n\hat{\theta}_i^0} \tag{2.5}$$

The asymptotic distribution of Q then may depend on the method employed for estimation. When the estimates are found by the method of maximum likelihood for the grouped data, the $L(\hat{\omega})$ in the likelihood-ratio test statistic is $L(\hat{\theta}_1^0, \hat{\theta}_2^0, \ldots, \hat{\theta}_k^0)$, where the $\hat{\theta}_i^0$ are the MLEs of the θ_i^0 under H_0. The derivation of the distribution of T and therefore Q goes through exactly as before except that the dimension of the space ω is increased. The degrees of freedom for Q then are $k - 1 - s$, where s is the number of independent parameters in $F_0(x)$ which had to be estimated from the grouped data in order to estimate all the θ_i^0. In the normal goodness-of-fit test, for example, the μ and σ parameter estimates would be calculated from the grouped data and used with tables of the normal distribution to find the $n\hat{\theta}_i^0$, and the degrees of freedom for k categories would be $k - 3$. When the original data are ungrouped and the MLEs are based on the likelihood function of all the observations, the theory is different. Chernoff and Lehmann (1954) have shown that the limiting distribution of Q is not the chi square in this case and that $P(Q > \chi_\alpha^2) > \alpha$. The test is then anticonservative. Their investigation showed that the error is considerably more serious for the normal distribution than the Poisson. A possible adjustment is discussed in their paper. In practice, however, the statistic in (2.5) is often treated as a chi-square variable anyway.

Example 2.1 A quality control engineer has taken 50 samples of size 13 each from a production process. The numbers of defectives for these samples are recorded below. Test the null hypothesis at level 0.05 that the number of defectives follows

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

(a) The Poisson distribution
(b) The binomial distribution

Number of defects	*Number of samples*
0	10
1	24
2	10
3	4
4	1
5	1
6 or more	0

Solution Since the data are grouped and both of the hypothesized null distributions are discrete, the chi-square goodness-of-fit test is appropriate. Since no parameters are specified, they must be estimated from the data in order to carry out the test in both (a) and (b).

Solution to (a) The Poisson distribution is $f(x) = e^{-\mu}\mu^x / x!$ for $x = 0, 1, 2, \ldots,$ where μ here is the mean number of defectives in a sample of size 13. The maximum-likelihood estimate of μ is the mean number of defectives in the 50 samples, that is,

$$\hat{\mu} = \frac{0(10) + 1(24) + 2(10) + 3(4) + 4(1) + 5(1)}{50} = \frac{65}{50} = 1.3$$

We use this value in $f(x)$ to estimate the probabilities as $\hat{\theta}_i$ and to compute $\hat{e}_i = 50\hat{\theta}_i$. The calculations are shown in Table 2.1. Notice that the final $\hat{\theta}$ is not for exactly 5 defects but rather for 5 or more; this is necessary to make $\sum \hat{\theta} = 1$. The final $\hat{e}$ is less than one, so it is combined with the category above before calculating Q. The final result is

Table 2.1 Calculation of *Q* for Example 2.1(a)

Defects	f	$\hat{\theta}$	$\hat{e}$	$(f - \hat{e})^2/\hat{e}$
0	10	0.2725	13.625	0.9644
1	24	0.3543	17.715	2.2298
2	10	0.2303	11.515	0.1993
3	4	0.0998	4.990	0.1964
4	1	0.0324	1.620 } 2.155	0.0111
5 or more	1	0.0107	0.535 }	
				3.6010

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

$Q = 3.6010$ with 3 degrees of freedom; we start out with $k - 1 = 5$ degree of freedom and lose one for estimating θ and one more for combining the last two categories. Table B shows the 0.05 critical value for the chi-square distribution with 3 degrees of freedom is 7.81. Our $Q = 3.6010$ is smaller than this value, so we cannot reject the null hypothesis. In terms of the P value, the approximate P value is the right-tail probability $P(Q \geqslant 3.601)$ where Q follows a chi square distribution with 3 degrees of freedom. Using EXCEL, for example, the P value is found as 0.3078. Note that using Table B, we could say that the P value is between 0.25 and 0.50. Thus, our conclusion about the Poisson distribution is that we cannot reject the null hypothesis.

Solution to (*b*) The null hypothesis is that the number of defectives in each sample of 13 follows the binomial distribution with $n = 13$ and p is the probability of a defective in any sample. The maximum-likelihood estimate of p is the total number of defectives, which we found in (a) to be 65, divided by the $50(13) = 650$ observations, or $p = 65/650 = 0.1$. This is the value we use in the binomial distribution (or Table C) to find $\hat{\theta}$ and $\hat{e} = 50\hat{\theta}$ in Table 2.2. The final result is $Q = 2.9680$, again with 3 degrees of freedom, so the critical value at the 0.05 level is again 7.81. The approximate P value using EXCEL is 0.3966. Our conclusion about the binomial distribution is that we cannot reject the null hypothesis.

This example illustrates a common result with chi-square goodness-of-fit tests, i.e., that each of two (or more) different null hypotheses may be accepted for the same data set. Obviously, the true distribution cannot be both binomial and Poisson at the same time. Thus, the appropriate conclusion on the basis of a chi-square goodness-of-fit test is that we do not have enough information to distinguish between these two distributions.

Table 2.2 Calculation of Q for Example 2.1(b)

Defects	f	$\hat{\theta}$		$\hat{e}$		$(f - \hat{e})^2/\hat{e}$
0	10	0.2542		12.710		0.5778
1	24	0.3671		18.355		1.7361
2	10	0.2448		12.240		0.4099
3	4	0.0997		4.986		0.1950
4	1	0.0277	} 0.0342	1.385	} 1.710	0.0492
5 or more	1	0.0065		0.325		
						2.9680

The STATXACT solutions to Example 2.1 are shown below. Note that the numerical value of Q in each case agrees with the hand calculations. Each printout shows the degrees of freedom as 4 instead of 3 because the computer did not know that the expected frequencies entered were calculated by estimating one parameter from the data in each case. The P values do not agree because the degrees of freedom are different.

```
****************************************
STATXACT SOLUTION TO EXAMPLE 2.1(a)
****************************************
CHI-SQUARE GOODNESS OF FIT TEST

    Statistic based on the observed 5 categories :
  CH(x) = Pearson Chi-Square Statistic =        3.601

Asymptotic p-value:(based on Chi-Square distribution with 4 df )
        Pr { CH(X) .GE.        3.601 } =       0.4627
```

```
****************************************
STATXACT SOLUTION TO EXAMPLE 2.1(b)
****************************************
CHI-SQUARE GOODNESS OF FIT TEST

    Statistic based on the observed 5 categories :
  CH(x) = Pearson Chi-Square Statistic =        2.968

Asymptotic p-value:(based on Chi-Square distribution with 4 df )
        Pr { CH(X) .GE.        2.968 } =       0.5633
```

4.3 THE KOLMOGOROV-SMIRNOV ONE-SAMPLE STATISTIC

In the chi-square goodness-of-fit test, the comparison between observed and expected class frequencies is made for a set of k groups. Only k comparisons are made even though there are n observations, where $k \leqslant n$. If the n sample observations are values of a continuous random variable, as opposed to strictly categorical data, comparisons can be made between observed and expected cumulative relative frequencies for each of the different observed values. The cumulative distribution function of the sample or the *empirical distribution function* defined in Section 2.3 is an estimate of the population cdf. Several goodness-of-fit test statistics are functions of the deviations between the observed cumulative distribution and the corresponding cumulative probabilities expected under the null hypothesis. The function of these deviations used to perform a test might be the sum of squares, or absolute values, or the maximum deviation, to name only a few. The best-known test is the Kolmogorov-Smirnov one-sample statistic,which will be covered in this section.

The Kolomogorov-Smirnov one-sample statistic is based on the differences between the hypothesized cumulative distribution function $F_0(x)$ and the empirical distribution function of the sample $S_n(x)$ for all x. The empirical distribution function of the sample was defined in Chapter 2 as $S_n(x)$, the proportion of sample observations that are less than or equal of x for all real numbers x. We showed there that $S_n(x)$ provides a consistent point estimator for the true distribution $F_x(x)$. Further, by the Glivenko-Cantelli Theorem in Chapter 2, we know that as n increases, the step function $S_n(x)$, with jumps occurring at the values of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ for the sample, approaches the true distribution function $F_X(x)$ for all x. Therefore, for large n, the deviations between the true function and its statistical image, $|S_n(x) - F_X(x)|$, should be small for all values for x. This result suggests that the statistic

$$D_n = \sup_x |S_n(x) - F_X(x)| \tag{3.1}$$

is, for any n, a reasonable measure of the accuracy of our estimate.

This D_n statistic, called the *Kolmogorov-Smirnov one-sample statistic*, is particularly useful in nonparametric statistical inference because the probability distribution of D_n does not depend on $F_X(x)$ as long as F_X is continuous. Therefore, D_n is called a distribution-free statistic.

The directional deviations defined as

$$D_n^+ = \sup_x [S_n(x) - F_X(x)] \qquad D_n^- = \sup_x [F_X(x) - S_n(x)] \tag{3.2}$$

are called the *one-sided Kolmogorov-Smirnov statistics*. These measures are also distribution free, as proved in the following theorem.

Theorem 3.1 *The statistics D_n, D_n^+ and D_n^- are completely distribution free for any continuous F_X.*

Proof $D_n = \sup_x |S_n(x) - F_X(x)| = \max_x (D_n^+, D_n^-)$

Defining the additional order statistics $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$, we can write

$$S_n(x) = \frac{i}{n} \qquad \text{for } X_{(i)} \leqslant x < X_{(i+1)}, \; i = 0, 1, \ldots, n$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Therefore, we have

$$\begin{aligned}
D_n^+ &= \sup_x [S_n(x) - F_X(x)] = \max_{0 \leqslant i \leqslant n} \sup_{X_{(i)} \leqslant x < X_{(i+1)}} [S_n(x) - F_X(x)] \\
&= \max_{0 \leqslant i \leqslant n} \sup_{X_{(i)} \leqslant x < X_{(i+1)}} \left[\frac{i}{n} - F_X(x)\right] \\
&= \max_{0 \leqslant i \leqslant n} \left[\frac{i}{n} - \inf_{X_{(i)} \leqslant x < X_{(i+1)}} F_X(x)\right] \\
&= \max_{0 \leqslant i \leqslant n} \left[\frac{i}{n} - F_X(X_{(i)})\right] \\
&= \max\left\{\max_{1 \leqslant i \leqslant n} \left[\frac{i}{n} - F_X(X_{(i)})\right], 0\right\}
\end{aligned} \tag{3.3}$$

Similarly

$$\begin{aligned}
D_n^- &= \max\left\{\max_{1 \leqslant i \leqslant n} \left[F_X(X_{(i)}) - \frac{i-1}{n}\right], 0\right\} \\
D_n &= \max\left\{\max_{1 \leqslant i \leqslant n} \left[\frac{i}{n} - F_X(X_{(i)})\right], \max_{1 \leqslant i \leqslant n} \left[F_X(X_{(i)}) - \frac{i-1}{n}\right], 0\right\}
\end{aligned} \tag{3.4}$$

The probability distributions of D_n, D_n^+, and D_n^- therefore depend only on the random variables $F_X(X_{(i)}), i = 1, 2, \ldots, n$. These are the order statistics from the uniform distribution on (0,1), regardless of the original F_X as long as it is continuous, because of the probability integral transformation discussed in Chapter 2. Thus D_n, D_n^+ and D_n^- have distributions which are independent of the particular F_X.

A simpler proof can be given by making the transformation $u = F_X(x)$ in D_n, D_n^+ or D_n^-. This will be left to the reader as an exercise. The above proof has the advantage of giving definitions of the Kolmogorov-Smirnov statistics in terms of order statistics.

In order to use the Kolmogorov statistics for inference, their sampling distributions must be known. Since the distributions are independent of F_X, we can assume without loss of generality that F_X is the uniform distribution on (0,1). The derivation of the distribution of D_n is rather tedious. However, the approach below illustrates a number of properties of order statistics and is therefore included here. For an interesting alternative derivation, see Massey (1950).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Theorem 3.2 *For* $D_n = \sup_x |S_n(x) - F_X(x)|$, *where* $F_X(x)$ *is any continuous cdf, we have*

$$P\left(D_n < \frac{1}{2n} + \nu\right) = \begin{cases} 0 & \text{for } \nu \leqslant 0 \\ \int_{1/2n-\nu}^{1/2n+\nu} \int_{1/3n-\nu}^{1/3n+\nu} \cdots \int_{(2n-1)/2n-\nu}^{(2n-1)/2n+\nu} & \\ \quad \times f(u_1, u_2, \ldots, u_n)\, du_n \cdots du_1 & \text{for } 0 < \nu < \frac{2n-1}{2n} \\ 1 & \text{for } \nu \geqslant \frac{2n-1}{2n} \end{cases}$$

where

$$f(u_1, u_2, \ldots, u_n) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < \cdots < u_n < 1 \\ 0 & \text{otherwise} \end{cases}.$$

Proof As explained above, $F_X(x)$ can be assumed to be the uniform distribution on (0,1). We shall first determine the relevant domain of ν. Since both $S_n(x)$ and $F_X(x)$ are between 0 and $1, 0 \leqslant D_n \leqslant 1$ always. Therefore we must determine $P(D_n < c)$ only for $0 < c < 1$, which here requires

$$0 < \frac{1}{2n} + \nu < 1 \qquad \text{or} \qquad -\frac{1}{2n} < \nu < \frac{2n-1}{2n}$$

Now, for all $-1/2n < \nu < (2n-1)/2n$, where $X_{(0)} = 0$ and $X_{(n+1)} = 1$,

$$\begin{aligned} p\left(D_n < \frac{1}{2n} + \nu\right) &= P\left[\sup_x |S_n(x) - x| < \frac{1}{2n} + \nu\right] \\ &= \left[|S_n(x) - x| < \frac{1}{2n} + \nu, \text{for all } x\right] \\ &= P\left[\left|\frac{i}{n} - x\right| < \frac{1}{2n} + \nu \text{ for } X_{(i)} \leqslant x < X_{(i+1)},\right. \\ &\qquad \left.\text{for all } i = 0, 1, \ldots, n\right] \\ &= P\left[\frac{i}{n} - \frac{1}{2n} - \nu < x < \frac{i}{n} + \frac{1}{2n} + \nu,\right. \\ &\qquad \left.\text{for all } i = 0, 1, \ldots, n\right] \\ &= P\left[\frac{2i-1}{2n} - \nu < x < \frac{2i+1}{2n} + \nu,\right. \\ &\qquad \left.\text{for } X_{(i)} \leqslant x < X_{(i+1)}, \text{for all } i = 0, 1, \ldots, n\right] \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Consider any two consecutive values of i. We must have, for any $0 \leqslant i \leqslant n-1$, both

$$A_i\colon \left\{\frac{2i-1}{2n} - \nu < x < \frac{2i+1}{2n} + \nu \text{ for } X_{(i)} \leqslant x \leqslant X_{(i+1)}\right\}$$

and

$$A_{i+1}\colon \left\{\frac{2i-1}{2n} - \nu < x < \frac{2i+3}{2n} + \nu \text{ for } X_{(i+1)} \leqslant x \leqslant X_{(i+2)}\right\}$$

Since $X_{(i+1)}$ is the random variable common to both events and the common set of x is $(2i+1)/2n - \nu < x < (2i+1)/2n + \nu$ for $\nu \geqslant 0$, the event $A_i \cap A_{i+1}$ for any $0 \leqslant i \leqslant n-1$ is

$$\frac{2i-1}{2n} - \nu < X_{(i+1)} < \frac{2i+1}{2n} + \nu \qquad \text{for all } \nu \geqslant 0$$

In other words,

$$\frac{2i-1}{2n} - \nu < x < \frac{2i+1}{2n} + \nu \qquad \begin{array}{l}\text{for } X_{(i)} \leqslant x \leqslant X_{(i+1)} \\ \text{for all } i = 0, 1, \ldots, n\end{array}$$

if and only if

$$\frac{2i+1}{2n} - \nu < X_{(i+1)} < \frac{2i+1}{2n} + \nu \qquad \begin{array}{l}\text{for all } i = 0, 1, \ldots, n-1 \\ \nu \geqslant 0\end{array}$$

The joint probability distribution of the order statistics is

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n) = n! \qquad \text{for } \ x_1 < x_2 < \cdots < x_n < 1$$

Putting all this together now, we have

$$\begin{aligned}
&P\left(D_n < \frac{1}{2n} + \nu\right) \qquad \text{for all} \quad -\frac{1}{2n} < \nu < \frac{2n-1}{2n} \\
&= P\left(\frac{2i+1}{2n} - \nu < X_{(i+1)} < \frac{2i+1}{2n} + \nu \quad \text{for all} \ \ i = 0, 1, \ldots, n-1\right) \\
&\qquad \text{for all } 0 \leqslant \nu < \frac{2n-1}{2n} \\
&= P\left[\left(\frac{1}{2n} - \nu < X_{(1)} < \frac{1}{2n} + \nu\right) \bigcap \left(\frac{3}{2n} - \nu < X_{(2)} < \frac{3}{2n} + \nu\right)\right. \\
&\qquad \left. \times \bigcap \cdots \bigcap \left(\frac{2n-1}{2n} - \nu < X_{(n)} < \frac{2n-1}{2n} + \nu\right)\right] \\
&\qquad\quad \text{for all} \quad 0 \leqslant \nu < \frac{2n-1}{2n}
\end{aligned}$$

which is equivalent to the stated integral.

This result is tedious to evaluate as it must be used with care. For the sake of illustration, consider $n = 2$. For all $0 \leqslant i \leqslant 3/4$,

$$P(D_2 < 1/4 + \nu) = 2! \int_{1/4-\nu}^{1/4+\nu} \int_{3/4-\nu}^{3/4+\nu} du_2\, du_1$$
$$0 < u_1 < u_2 < 1$$

The limits overlap when $1/4 + \nu \geqslant 3/4 - \nu$, or $\nu \geqslant 1/4$ When $0 \leqslant \nu < 1/4$, we have $u_1 < u_2$ automatically. Therefore, for $0 \leqslant \nu < 1/4$,

$$P(D_2 < 1/4 + \nu) = 2 \int_{1/4-\nu}^{1/4+\nu} \int_{3/4-\nu}^{3/4+\nu} du_2\, du_1 = 2(2\nu)^2$$

But for $1/4 \leqslant \nu \leqslant 3/4$, the region of integration is as illustrated in Figure 3.1. Dividing the integral into two pieces, we have for $1/4 \leqslant \nu < 3/4$,

$$P(D_2 < 1/4 + \nu) = 2\left[\int_{3/4-\nu}^{1/4+\nu} \int_{u_1}^{1} du_2\, du_1 + \int_{0}^{3/4-\nu} \int_{3/4-\nu}^{1} du_2\, du_1\right]$$
$$= -2\nu^2 + 3\nu - 1/8$$

Collecting the results for all ν,

$$P(D_2 < 1/4 + \nu) = \begin{cases} 0 & \text{for } \nu \leqslant 0 \\ 2(2\nu)^2 & \text{for } 0 < \nu < 1/4 \\ -2\nu^2 + 3\nu - 0.125 & \text{for } 1/4 \leqslant \nu < 3/4 \\ 1 & \text{for } \nu \geqslant 3/4 \end{cases}$$

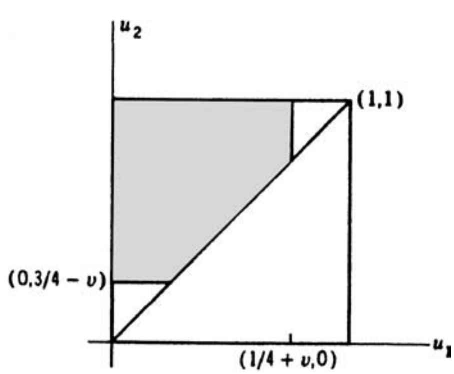


Fig. 3.1 Shaded area is region of integration for $n = 2$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For any given ν and n, we can evaluate $P(D_n < 1/2n + \nu)$ or use Table 1 of Birnbaum (1952). The inverse procedure is to find that number $D_{n,\alpha}$ such that $P(D_n > D_{n,\alpha}) = \alpha$. In our numerical example with $n = 2$, $\alpha = 0.05$, we find ν such that

$$P(D_2 > 1/4 + \nu) = 0.05 \qquad \text{or} \qquad P(D_2 < 1/4 + \nu) = 0.95$$

and then set $D_{2,0.05} = 1/4 + \nu$. From the previous evaluation of the D_2 sampling distribution, either

$$2(2\nu)^2 = 0.95 \qquad \text{and} \qquad 0 < \nu < 1/4$$

or

$$-2\nu^2 + 3\nu - 0.125 = 0.95 \qquad \text{and} \qquad 1/4 \leqslant \nu < 3/4$$

The first result has no solution, but the second yields the solution $\nu = 0.5919$. Therefore, $D_{2,0.05} = 0.8419$.

Numerical values of $D_{n,\alpha}$ are given in Table F of the Appendix for $n \leqslant 40$ and selected tail probabilities α, and approximate values are given for larger n. More extensive tables are given in Dunstan, Nix, and Reynolds (1979, p. 42) for $n \leqslant 100$.

For large samples, Kolmogorov (1933) derived the following convenient approximation to the sampling distribution of D_n, and Smirnov (1939) gave a simpler proof. The result is given here without proof.

Theorem 3.3 *If F_X is any continuous distribution function, then for every $d > 0$,*

$$\lim_{n\to\infty} P\left(D_n \leqslant d/\sqrt{n}\right) = L(d)$$

where

$$L(d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2d^2}$$

The function $L(d)$ has been tabulated in Smirnov (1948). Some of the results for the asymptotic approximation to $D_{n,\alpha} = d_\alpha/\sqrt{n}$ are:

$P(D_n > d_\alpha/\sqrt{n})$	0.20	0.15	0.10	0.05	0.01
d_α	1.07	1.14	1.22	1.36	1.63

The approximation has been found to be close enough for practical application as long as n exceeds 35. A comparison of exact and asymptotic values of $D_{n,\alpha}$ for $\alpha = 0.01$ and 0.05 is given in Table 3.1.

Since the one-sided Kolmogorov-Smirnov statistics are also distribution-free, knowledge of their sampling distributions would make them useful in nonparametric statistical inference as well. Their exact sampling distributions are considerably easier to derive than that for D_n. Only the statistic D_n^+ is considered in the following theorem, but D_n^+ and D_n^- have identical distributions because of symmetry.

Theorem 3.4 *For* $D_n^+ = \sup_x [S_n(x) - F_X(x)]$ *where* $F_X(x)$ *is any continuous cdf, we have*

$$P(D_n^+ < c) = \begin{cases} 0 & c \leqslant 0 \\ \int_{1-c}^{1} \int_{(n-1)/n-c}^{u_n} \cdots \int_{2/n-c}^{u_3} \int_{1/n-c}^{u_2} f(u_1, u_2, \ldots, u_n)\, du_1 \cdots du_n & 0 < c < 1 \\ 1 & c \geqslant 1 \end{cases}$$

where

$$f(u_1, u_2, \ldots, u_n) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < \cdots < u_n < 1 \\ 0 & \text{otherwise} \end{cases}$$

Proof As before, we first assume without loss of generality that F_X is the uniform distribution on (0,1). Then we can write

Table 3.1 Exact and asymptotic values of $D_{n,\alpha}$ such that $P(D_n > D_{n,\alpha}) = \alpha$ for $\alpha = 0.01, 0.05$

	Exact		*Asymptotic*		*Ratio A/E*	
n	*0.05*	*0.01*	*0.05*	*0.01*	*0.05*	*0.01*
2	0.8419	0.9293	0.9612	1.1509	1.142	1.238
3	0.7076	0.8290	0.7841	0.9397	1.108	1.134
4	0.6239	0.7341	0.6791	0.8138	1.088	1.109
5	0.5633	0.6685	0.6074	0.7279	1.078	1.089
10	0.4087	0.4864	0.4295	0.5147	1.051	1.058
20	0.2939	0.3524	0.3037	0.3639	1.033	1.033
30	0.2417	0.2898	0.2480	0.2972	1.026	1.025
40	0.2101	0.2521	0.2147	0.2574	1.022	1.021
50	0.1884	0.2260	0.1921	0.2302	1.019	1.018

Source: Z. W. Birnbaum (1952): Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size, *J. Am. Statist. Assoc.*, **47**, 431, Table 2.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$D_n^+ = \max\left[\max_{1\leqslant i\leqslant n}\left(\frac{i}{n} - X_{(i)}\right),\ 0\right]$$

the form found in (3.3). For all $0 < c < 1$, we have

$$\begin{aligned}
P(D_n^+ < c) &= P\left[\max_{1\leqslant i\leqslant n}\left(\frac{1}{n} - X_{(i)}\right) < c\right] \\
&= P\left(\frac{i}{n} - X_{(i)} < c \text{ for all } i = 1, 2, \ldots, n\right) \\
&= P\left(X_{(i)} > \frac{i}{n} - c \text{ for all } i = 1, 2, \ldots, n\right) \\
&= \int_{1-c}^{\infty}\int_{(n-1)/n-c}^{\infty}\cdots\int_{2/n-c}^{\infty}\int_{1/n-c}^{\infty} f(x_1, x_2, \ldots, x_n)\,dx_1\cdots dx_n
\end{aligned}$$

where

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} n! & \text{for } 0 < x_1 < x_2 < \cdots < x_n < 1 \\ 0 & \text{otherwise} \end{cases}$$

which is equivalent to the stated integral.

Another form of this result, due to Birnbaum and Tingey (1951), which is more computationally tractable, is

$$P(D_n^+ > c) = (1-c)^n + c\sum_{j=1}^{[n(1-c)]}\binom{n}{j}\left(1 - c - \frac{j}{n}\right)^{n-j}\left(c + \frac{j}{n}\right)^{j-1} \tag{3.5}$$

The equivalence of the two forms can be shown by induction. Birnbaum and Tingey give a table of those values $D_{n,\alpha}^+$, which satisfy $P(D_n > D_{n,\alpha}^+) = \alpha$ for $\alpha = 0.01, 0.05, 0.10$ and selected values of n. For large samples, we have the following theorem, which is given here without proof.

Theorem 3.5 *If F_X is any continuous distribution function, then for every $d \geqslant 0$,*

$$\lim_{n\to\infty} P\left(D_n^+ < d/\sqrt{n}\right) = 1 - e^{-2d^2}$$

As a result of this theorem, chi-square tables can be used for the distribution of a function of D_n^+ because of the following corollary.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Corollary 3.5 *If F_X is any continuous distribution function, then for every $d \geqslant 0$, the limiting distribution of $V = 4nD_n^{+2}$, as $n \to \infty$, is the chi-square distribution with 2 degrees of freedom.*

Proof We have $D_n^+ < d/\sqrt{n}$ if and only if $4nD_n^{+2} < 4d^2$ or $V < 4d^2$. Therefore

$$\lim_{n\to\infty} P(V < 4d^2) = \lim_{n\to\infty} P\left(D_n^+ < d/\sqrt{n}\right) = 1 - e^{-2d^2} = 1 - e^{-4d^2/2}$$

$$\lim_{n\to\infty} P(V < c) = 1 - e^{-c/2} \qquad \text{for all } c > 0$$

The right-hand side is the cdf of a chi-square distribution with 2 degrees of freedom.

As a numerical example of how this corollary enables us to approximate $D_{n,\alpha}^+$, let $\alpha = 0.05$. Table B of the Appendix shows that 5.99 is the 0.05 critical point of chi square with 2 degrees of freedom. The procedure is to set $4nD_{n,0.05}^{+2} = 5.99$ and solve to obtain

$$D_{n,0.05}^+ \sqrt{1.4975/n} = 1.22/\sqrt{n}$$

4.4 APPLICATIONS OF THE KOLMOGOROV-SMIRNOV ONE-SAMPLE STATISTICS

The statistical use of the Kolmogorov-Smirnov statistic in a goodness-of-fit type of problem is obvious. Assume we have the random sample $X_1, X_2, \ldots, X_n$ and the hypothesis-testing situation H_0: $F_X(x) = F_0(x)$ for all x, where $F_0(x)$ is a completely specified continuous distribution function.

Since $S_n(x)$ is the statistical image of the population distribution $F_X(x)$, the differences between $S_n(x)$ and $F_0(x)$ should be small for all x except for sampling variation, if the null hypothesis is true. For the usual two-sided goodness-of-fit alternative.

$$H_1\colon F_X(x) \neq F_0(x) \qquad \text{for some } x$$

large absolute values of these deviations tend to discredit the hypothesis. Therefore, the Kolmogorov-Smirnov goodness-of-fit test with significance level α is to reject H_0 when $D_n > D_{n,\alpha}$. From the Glivenko-Cantelli theorem of Chapter 2, we know that $S_n(x)$ converges to $F_X(x)$ with probability 1, which implies consistency.

The value of the Kolmogorov-Smirnov goodness-of-fit statistic D_n in (3.1) can be calculated using (3.4) if all n observations have different numerical values (no ties). However, the expression below is considerably easier for algebraic calculation and applies when ties are present. The formula is

$$D_n = \sup_x |S_n(x) - F_0(x)| = \max_x [|S_n(x) - F_0(x)|, |S_n(x - \varepsilon) - F_0(x)|]$$

where ε denotes any small positive number. Example 4.1 will illustrate this easy algebraic method of calculating D_n. Quantiles of the exact null distribution of D_n are given in Table F in the Appendix for $n \leqslant 40$, along with approximate values for $n > 40$. The appropriate critical region is for D_n large.

Example 4.1 The 20 observations below were chosen randomly from the continuous uniform distribution over (0,1), recorded to four significant figures, and rearranged in increasing order of magnitude. Determine the value of D_n, and test the null hypothesis that the square roots of these numbers also have the continuous uniform distribution, over (0,1).

0.0123	0.1039	0.1954	0.2621	0.2802
0.3217	0.3645	0.3919	0.4240	0.4814
0.5139	0.5846	0.6275	0.6541	0.6889
0.7621	0.8320	0.8871	0.9249	0.9634

Solution The calculations needed to find D_n are shown in Table 4.1. The entries in the first column, labeled x, are not the observations above, but their respective square roots, because the null hypothesis is concerned with the distribution of these square roots. The $S_n(x)$ are the proportions of observed values less than or equal to each different observed x. The hypothesized distribution here is $F_0(x) = x$, so the third column is exactly the same as the first column. The fourth column is the difference $S_n(x) - F_0(x)$. The fifth column is the difference $S_n(x - \varepsilon) - F_0(x)$, that is, the difference between the S_n value for a number slightly smaller than an observed x and the F_0 value for that observed x. Finally the sixth and seventh columns are the absolute values of the differences of the numbers in the fourth and fifth columns. The supremum is the largest entry in either of the last two columns; its value here is $D_n = 0.36$. Table F shows that the 0.01 level rejection region for $n = 20$ is $D_n \geqslant 0.352$, so we reject the null hypothesis that these numbers are uniformly distributed.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 4.1 Calculation of D_n for Example 4.1

x	$S_n(x)$	$F_0(x)$	$S_n(x)-F_0(x)$	$S_n(x-\varepsilon)-F_0(x)$	$\|S_n(x)-F_0(x)\|$	$\|S_n(x-\varepsilon)-F_0(x)\|$
0.11	0.05	0.11	−0.06	−0.11	0.06	0.11
0.32	0.10	0.32	−0.22	−0.27	0.22	0.27
0.44	0.15	0.44	−0.29	−0.34	0.29	0.34
0.51	0.20	0.51	−0.31	−0.36	0.31	0.36
0.53	0.25	0.53	−0.28	−0.33	0.28	0.33
0.57	0.30	0.57	−0.27	−0.32	0.27	0.32
0.60	0.35	0.60	−0.25	−0.30	0.25	0.30
0.63	0.40	0.63	−0.23	−0.28	0.23	0.28
0.65	0.45	0.65	−0.20	−0.25	0.20	0.25
0.69	0.50	0.69	−0.19	−0.24	0.19	0.24
0.72	0.55	0.72	−0.17	−0.22	0.17	0.22
0.76	0.60	0.76	−0.16	−0.21	0.16	0.21
0.79	0.65	0.79	−0.14	−0.19	0.14	0.19
0.81	0.70	0.81	−0.11	−0.16	0.11	0.16
0.83	0.75	0.83	−0.08	−0.13	0.08	0.13
0.87	0.80	0.87	−0.07	−0.12	0.07	0.12
0.91	0.85	0.91	−0.06	−0.11	0.06	0.11
0.94	0.90	0.94	−0.04	−0.09	0.04	0.09
0.96	0.95	0.96	−0.01	−0.06	0.01	0.06
0.98	1.00	0.98	0.02	−0.03	0.02	0.03

The theoretical justification behind this example is as follows. Let Y have the continuous uniform distribution on (0,1) so that $f_Y(y) = 1$ for $0 \leqslant y \leqslant 1$. Then the pdf of $X = \sqrt{Y}$ can be shown to be (the reader should verify this) $f_X(x) = 2x$ for $0 \leqslant x \leqslant 1$, which is not uniform. In fact, this is a beta distribution with parameters $a = 2$ and $b = 1$.

ONE-SIDED TESTS

With the statistics D_n^+ and D_n^-, it is possible to use Kolmogorov-Smirnov statistics for a one-sided goodness-of-fit test which would detect directional differences between $S_n(x)$ and $F_0(x)$. For the alternative

$$H_{1,+}\colon F_X(x) \geqslant F_0(x) \qquad \text{for all } x$$

the appropriate rejection region is $D_n^+ > D_{n,\alpha}^+$, and for the alternative

$$H_{1,-}\colon F_X(x) \leqslant F_0(x) \qquad \text{for all } x$$

H_0 is rejected when $D_n^- > D_{n,\alpha}^-$. Both of these tests are consistent against their respective alternatives.

Most tests of goodness of fit are two-sided and so applications of the one-sided Kolmogorov-Smirnov statistics will not be demonstrated here in detail. However, it is useful to know that the tail probabilities for the one-sided statistics are approximately one-half of the corresponding tail probabilities for the two-sided statistic. Therefore, approximate results can be obtained for the one-sided statistics by using the entries in Table F with each quantile being one-half of the value labeled. In our example, we find $D_n^+ = 0.02$ as the largest entry in the fourth column of Table 4.1. Now from Table F we see that for $n = 20$, the smallest critical value corresponding to a two-tailed P value of 0.200 is 0.232. Since the observed value 0.02 is smaller than 0.232, we conclude that the approximate P value for testing H_0 against $H_{1,+}$: $F_X(x) \geqslant F_0(x)$ for all x is larger than 0.100 so we fail to reject H_0 in favor of $H_{1,+}$. For the alternative $H_{1,-}$: $F_X(x) \leqslant F_0(x)$ we find $D_n^- = 0.36$ from the fifth column of Table 4.1. The approximate P value from Table F is $P < 0.005$ and so we reject H_0 in favor of $H_{1,-}$. If, for example, we observed $D_n^+ = 0.30$, the approximate P value is between 0.01 and 0.025.

The STATXACT solution to Example 4.1 is shown below. The values of all three of the D_n statistics and P values agree with ours. The STATXACT package also provides Kolmogorov-Smirnov type tests of goodness of fit for some specific distributions such as the binomial, Poisson, uniform, exponential and normal. Some of these will be discussed and illustrated later in this chapter.

```
************************************
STATXACT SOLUTION TO EXAMPLE 4.1
************************************

KOLMOGOROV-SMIRNOV ONE-SAMPLE TEST

Hypothesized distribution F(X)  : Uniform(Cont.);
                                  Min =      0.0000 Max =       1.000

Let S(X) be the empirical distribution.
                         Sample size : 20
Inference :
                                            Statistic
Item                     Sup{|S(X) - F(X)|}  Sup{S(X) - F(X)}  Sup{F(X) - S(X)}
Observed Statistic             0.3600            0.02000            0.3600
Stand. Statistic                1.610           -0.08944             1.610
Asymptotic p-value             0.0112             0.9841            0.0056
Exact p-value                  0.0079             0.9709            0.0040
Exact Point Prob.              0.0000             0.0000            0.0000
```

MINITAB offers the Kolmogorov-Smirnov test for the normal distribution and this will be discussed later in this chapter.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Two other useful applications of the Kolmogorov-Smirnov statistics relate to point and interval estimations of the unknown cdf F_X.

CONFIDENCE BANDS

One important statistical use of the D_n statistic is in finding confidence bands on $F_X(x)$ for all x. From Table F in the Appendix, we can find the number $D_{n,\alpha}$ such that

$$P(D_n > D_{n,\alpha}) = \alpha$$

This is equivalent to the statement

$$P[\sup_x |S_n(x) - F_X(x)| < D_{n,\alpha}] = 1 - \alpha$$

which means that

$$P[S_n(x) - D_{n,\alpha} < F_X(x) < S_n(x) + D_{n,\alpha} \text{ for all } x] = 1 - \alpha$$

But we know that $0 \leqslant F_X(x) \leqslant 1$ for all x, whereas the inequality in this probability statement admits numbers outside this range. Thus we define

$$L_n(x) = \max[S_n(x) - D_{n,\alpha}, 0]$$

and

$$U_n(x) = \min[S_n(x) + D_{n,\alpha}, 1]$$

and call $L_n(x)$ a lower confidence band and $U_n(x)$ an upper confidence band for the cdf F_X, with associated confidence coefficient $1 - \alpha$.

The simplest procedure in application is to graph the observed $S_n(x)$ as a step function and draw parallel lines at a distance $D_{n,\alpha}$ in either direction, but always within the unit square. When $n > 40$, the value $D_{n,\alpha}$ can be determined from the asymptotic distribution. Of course, this confidence-band procedure can be used to perform a test of the hypothesis $F_X(x) = F_0(x)$, since $F_0(x)$ lies wholly within the limits $L_n(x)$ and $U_n(x)$ if and only if the hypothesis cannot be rejected at a significance level α.

Similar applications of the D_n^- or D_n^+ statistics are obvious.

One criticism of confidence bands is that they are too wide, particularly in the tails of the distribution, where the order statistics have a lot of variation. Keeping this in mind, other approaches to constructing a confidence band on the cdf have been considered in the literature. One general idea is to base bands on $\sup_x[w\{F(x)\}|S_n(x) - F(x)|]$, where $w(x)$ is a suitable weight function. Some authors have

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

also considered restricting the range of x to a finite interval. For example, Doksum (1977) used $\max_{a \leqslant S_n(x) \leqslant b} \frac{|S_n(x)-F(x)|}{[F(x)\{1-F(x)\}]^{1/2}}$, where a and b are constants with $0 \leqslant a < b \leqslant 1$, to set up a confidence band for $F_X(x)$. The resulting band is slightly wider in the middle but is much narrower in the tails.

SAS provides confidence bands for the cdf for various confidence levels. We present an example using the data in Example 4.1 and the module "Interactive data analysis." Figure 4.1 shows a screenshot, helping the reader find the procedure under SAS Version 8. The output is shown in Figure 4.2. Note the slider in the output, which allows one to change, for example, the confidence level interactively, and examine the effect on the bands. Note also that the output includes the K-S test statistic and the associated P value, together with the estimates of the population mean and the standard deviation. The K-S test for the normal distribution with unknown mean and standard deviation is called Lilliefors's test and is discussed later in Section 4.5. In Figure 4.3 we show the output using data from Example 5.1.

Next we consider an application of the K-S test statistic to determine sample size.

DETERMINATION OF SAMPLE SIZE

In the case of a point estimate, the D_n statistic enables us to determine the minimum sample size required to guarantee, with a certain probability $1 - \alpha$, that the error in the estimate never exceeds a fixed positive value c. This allows us to formulate the sample size determination problem as follows. We want to find the minimum value of n that satisfies

$$P(D_n < c) = 1 - \alpha$$

This is equivalent to saying

$$1 - P(D_n < c) = P(D_n > c) = \alpha$$

and therefore c equals the value of $D_{n,\alpha}$ given in Table F of the Appendix. This means that the value of n can be read directly from Table F as that sample size corresponding to $D_{n,\alpha} = c$. If no $n \leqslant 40$ will meet the specified accuracy, the asymptotic distribution of Theorem 3.3 can be used by solving $c = d/\sqrt{n}$ for n, where $d/\sqrt{n}$ is given in the last row of Table F in the Appendix.

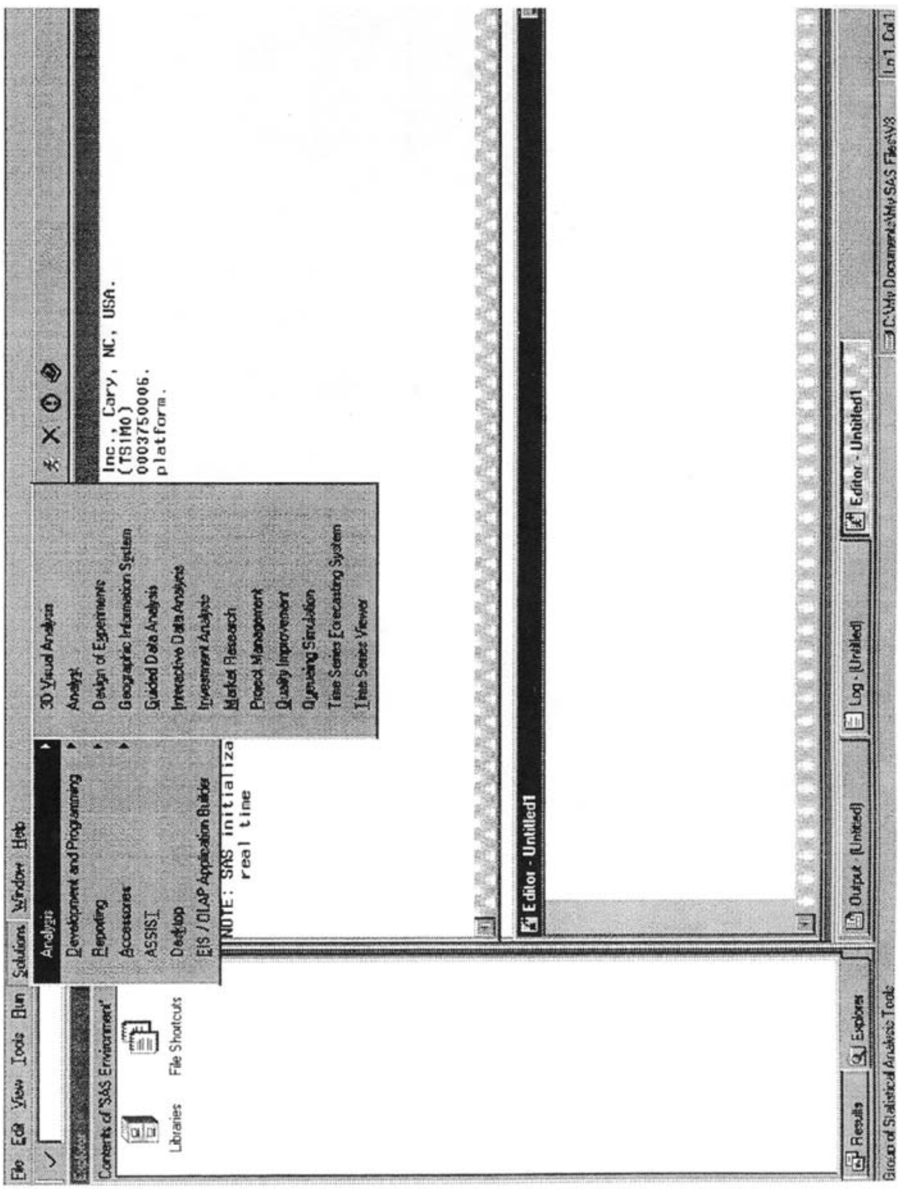


Fig. 4.1 SAS Interactive Data Analysis for Example 4.1.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

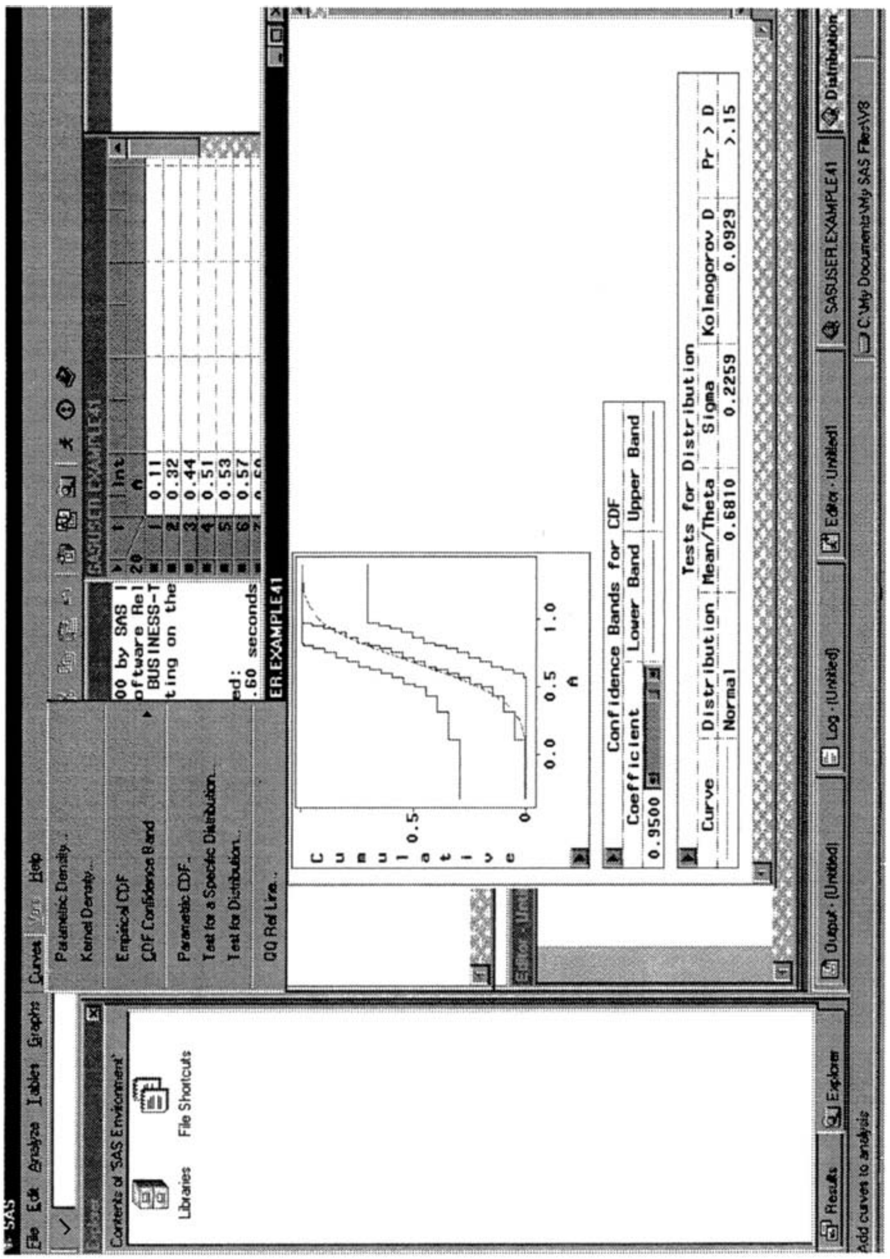


Fig. 4.2 SAS Interactive Data Analysis output for confidence band for Example 4.1.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

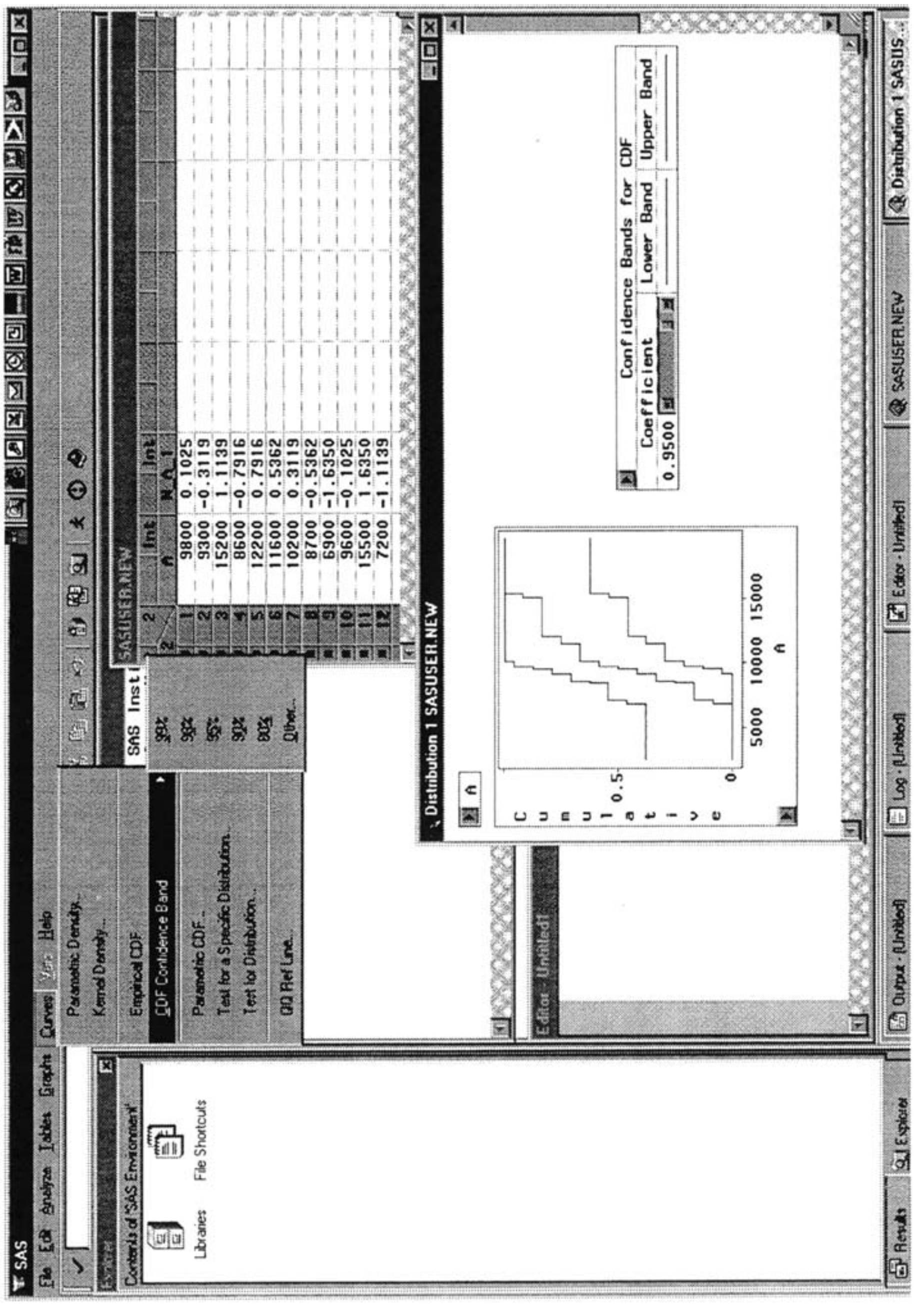


Fig. 4.3 SAS Interactive Data Analysis Confidence Bands output for Example 5.1.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For example, suppose we want to take a sample of size n and use the resulting $S_n(x)$ as a point estimate of $F_X(x)$ for each x. We want the error in this estimate to be no more than 0.25 with probability 0.98. How large a sample should be taken? We look down the $0.02 = 1 - 0.98$ column in Table F of the Appendix until we find the largest c that is less than or equal to 0.25. This entry is 0.247, which corresponds to a sample size of $n = 36$. If we want more precision in our point estimate and thereby specify a maximum error of 0.20 but keep the probability at 0.98, Table F shows that $n > 40$. The value is found by solving $1.52/\sqrt{n} = 0.20$ and we get $n = 57.76$, which is rounded up to require a sample size of 58 observations.

It should be noted that all of the theoretical properties of the Kolmogorov-Smirnov statistics required the assumption that F_X be continuous, since this is necessary to guarantee their distribution-free nature. The properties of the empirical distribution function given in Section 2.3, including the Glivenko-Cantelli theorem, do not require this continuity assumption. Furthermore, it is certainly desirable to have a goodness of fit test which can be used when the hypothesized distribution is discrete. Noether (1967, pp. 17–18) and others have shown that if the $D_{n,\alpha}$ values based on a continuous F_X are used in a discrete application, the same significance level is at most α. However, Slakter (1965) used Monte Carlo techniques to show that this procedure is extremely conservative. Conover (1972) found recursive relationships for the exact distribution of D_n for F_0 discrete.

Pettitt and Stephens (1977) give tables of exact tail probabilities of nD_n that can be used with F_0 discrete and also for grouped data from a continuous distribution as long as the expected frequencies are equal. They show that in these two cases, the test statistic can be written as

$$nD_n = \max_{1 \leqslant j \leqslant k} \left| \sum_{i=1}^{j} (F_i - e_i) \right|$$

because

$$S_n(x_j) = \sum_{i=1}^{j} \frac{F_i}{n} \qquad \text{and} \qquad F_0(x_j) = \sum_{i=1}^{j} \frac{e_i}{n}$$

for ordered $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$. This expression shows that the distribution of D_n depends on the chosen grouping and also shows explicitly the relationship between D_n and the chi-square statistic Q.

This exact test has greater power than the chi-square test in the case of grouped data.

4.5 LILLIEFORS'S TEST FOR NORMALITY

In this section we consider the problem of a goodness-of-fit test for the normal distribution with no specified mean and variance. This problem is very important in practice because the assumption of a general normal distribution with unknown μ and σ is necessary to so many classical statistical test and estimation procedures. In this case, note that the null hypothesis is composite because it states that the underlying distribution is some normal distribution. In general, K-S tests can be applied in the case of composite goodness-of-fit hypotheses after estimating the unknown parameters [$F_0(x)$ will then be replaced by $\hat{F}_0(x)$]. Unfortunately, the null distribution of the K-S test statistic with estimated parameters is far more complicated. This, of course, affects the P-value calculations. In the absence of any additional information, one approach could be to use the tables of the K-S test to approximate the P value or to find the approximate critical value. For the normal distribution, Lilliefors (1967) showed that using the usual critical points developed for the K-S test gives extremely conservative results. He then used Monte Carlo simulations to develop a table for the Kolmogorov-Smirnov statistic that gives accurate critical values. As before, the Kolmogorov-Smirnov two-sided statistic is defined as

$$D_n = \sup_x \left| S_n(x) - \hat{F}_0(x) \right|$$

Here $\hat{F}_0(x)$ is computed as the cumulative standard normal distribution $\Phi(z)$ where $z = (x - \bar{x})/s$ for each observed x, $\bar{x}$ is the mean of the sample of n observations, and s^2 is the unbiased estimator of σ^2 (computed with $n - 1$ in the denominator). The appropriate rejection region is in the right tail and Appendix Table O gives the exact tail probabilities computed by Monte Carlo simulations. This table is taken from Edgeman and Scott (1987), in which more samples were used to improve the accuracy of the original results given by Lilliefors (1967).

Example 5.1 A random sample of 12 persons is interviewed to estimate median annual gross income in a certain economically depressed town. Use the most appropriate test for the null hypothesis that income data below are normally distributed.

9,800	10,200	9,300	8,700	15,200	6,900
8,600	9,600	12,200	15,500	11,600	7,200

Solution Since the mean and variance are not specified, the most appropriate test is the Lilliefors's test. The first step is to calculate $\bar{x}$ and s. From the data we get $\sum x = 124,800$ and $\sum(x-\bar{x})^2 = 84{,}600{,}000$ so that $\bar{x} = 10{,}400$ and $s = \sqrt{84{,}600{,}000/11} = 2{,}773.25$. The corresponding standard normal variable is then $z = (x - 10{,}400)/2{,}773$. The calculations needed for D_n are shown in Table 5.1 (p. 132). We find $D_n = 0.1946$ and $P > 0.10$ for $n = 12$ from Table O. Thus, the null hypothesis that incomes are normally distributed is not rejected.

The following computer printouts illustrate the solution to Example 5.1 using the SAS and MINITAB packages.

```
*******************************************************
       SAS/ANALYST SOLUTION TO EXAMPLE 5.1
*******************************************************

              The UNIVARIATE Procedure
              Fitted Distribution for A

    Parameters for Normal Distribution

          Parameter   Symbol   Estimate
          Mean        Mu          10400
          Std Dev     Sigma    2773.249

Goodness-of-Fit Tests for Normal Distribution

Test                  ---Statistic----   -----p Value-----

Kolmogorov-Smirnov    D     0.19541250   Pr > D      >0.150
Cramer-von Mises      W-Sq  0.07074246   Pr > W-Sq   >0.250
Anderson-Darling      A-Sq  0.45877863   Pr > A-Sq    0.223
```

Note that both the MINITAB and SAS outputs refer to this as the K-S test and not the Lilliefors's test. Both calculate a modified K-S statistic using formulas given in D'Agostino and Stephens (1986); the results agree with ours to two decimal places. SAS also provides the results for two other tests, called the Anderson-Darling and the Cramér-von Mises tests (see Problem 4.14). In this particular example, each of the tests fails to reject the null hypothesis of normality. In MINITAB one can go from Stat to Basic Statistics to Normality Test and the output is a graph, called Normal Probability Plot. The grid on

```
*****************************************
     MINITAB SOLUTION TO EXAMPLE 5.1
*****************************************
```

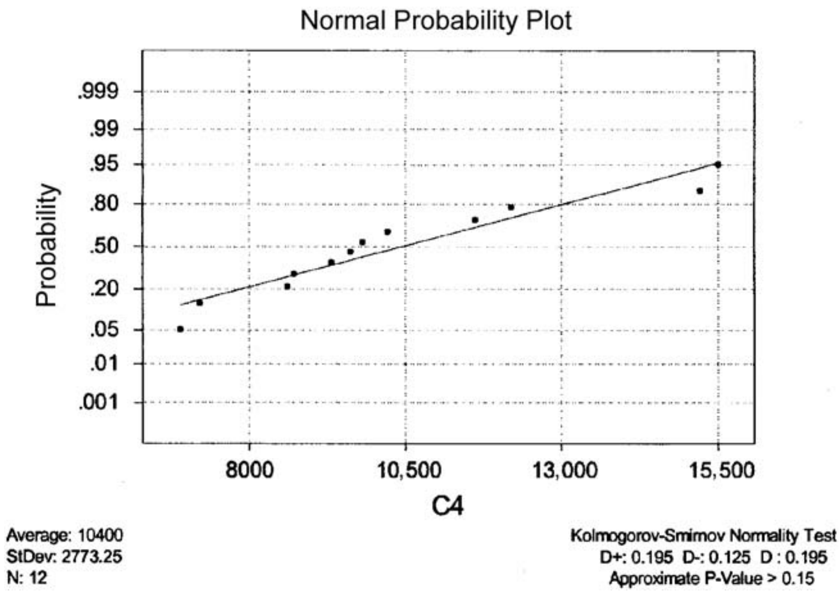


Table 5.1 Calculations for Lilliefors's statistic in Example 5.1

x	z	$S_n(x)$	$\Phi(z)$	$\lvert S_n(x) - \Phi(z)\rvert$	$\lvert S_n(x - \varepsilon) - \Phi(z)\rvert$
6900	−1.26	0.0833	0.1038	0.0205	0.1038
7200	−1.15	0.1667	0.1251	0.0416	0.0418
8600	−0.65	0.2500	0.2578	0.0078	0.0911
8700	−0.61	0.3333	0.2709	0.0624	0.0209
9300	−0.40	0.4167	0.3446	0.0721	0.0113
9600	−0.29	0.5000	0.3859	0.1141	0.0308
9800	−0.22	0.5833	0.4129	0.1704	0.0871
10200	−0.07	0.6667	0.4721	0.1946	0.1112
11600	0.43	0.7500	0.6664	0.0836	0.0003
12200	0.65	0.8333	0.7422	0.0911	0.0078
15200	1.73	0.9167	0.9582	0.0415	0.1249
15500	1.84	1.0000	0.9671	0.0329	0.0504
			1.0000		0

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

this graph resembles that found on a normal probability paper. The vertical axis is a probability scale and the horizontal axis is the usual data scale. The plotted points are the $S_n(x)$ values; the straight line is a least-squares line fitted to these points. If the points fall reasonably close to this straight line, the normal distribution hypothesis is supported. The P value given by both packages is an approximation based on linear interpolation in tables that are not the same as our Table O. See the documentation in the packages and D'Agostino and Stephens (1986) for additional details.

The SAS software package also provides a number of excellent (interactive) ways to study goodness of fit. We have given one illustration already for confidence bands using data from Example 4.1. Now we illustrate some of the other possibilities in Figures 5.1 through 5.6, using the data from Example 5.1 and the Interactive Data Analysis and the Analyst options, respectively, under SAS version 8.0. The highlights include a plot of the empirical cdf, a box plot, confidence bands, tests for a specific distribution the choices for which includes the normal, the lognormal, the exponential and the Weibull, a Q-Q plot together with a reference line and other available options. Also interesting is a slider (the bottom panel in the output shown) where one can "try out" various means and standard deviations to be compatible with the data on the basis of the K-S test. For details on the features, the reader is referred to SAS version 8.0 online documentations.

4.6 LILLIEFORS'S TEST FOR THE EXPONENTIAL DISTRIBUTION

Another important goodness-of-fit problem in practice is to test for the exponential distribution with no specified mean. This problem is important because the assumption of an exponential distribution with an unknown mean μ is made in many applications, particularly where the random variable under study is the waiting time, the time to the occurrence of an event. Lilliefors (1969) developed an analog of the Kolmogorov-Smirnov test in this situation and gave a table of critical values based on Monte Carlo simulations. As in the normal case with unknown parameters, the Kolmogorov-Smirnov two-sided statistic is defined as

$$D_n = \sup_x |S_n(x) - \hat{F}_0(x)|$$

Here $\hat{F}_0(x)$ is computed as $1 - e^{-x/\bar{x}} = \hat{F}_0(z) = 1 - e^{-z}$, say, where $\bar{x}$ is the sample mean and $z = x/\bar{x}$ for each observed x. Thus one forms

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

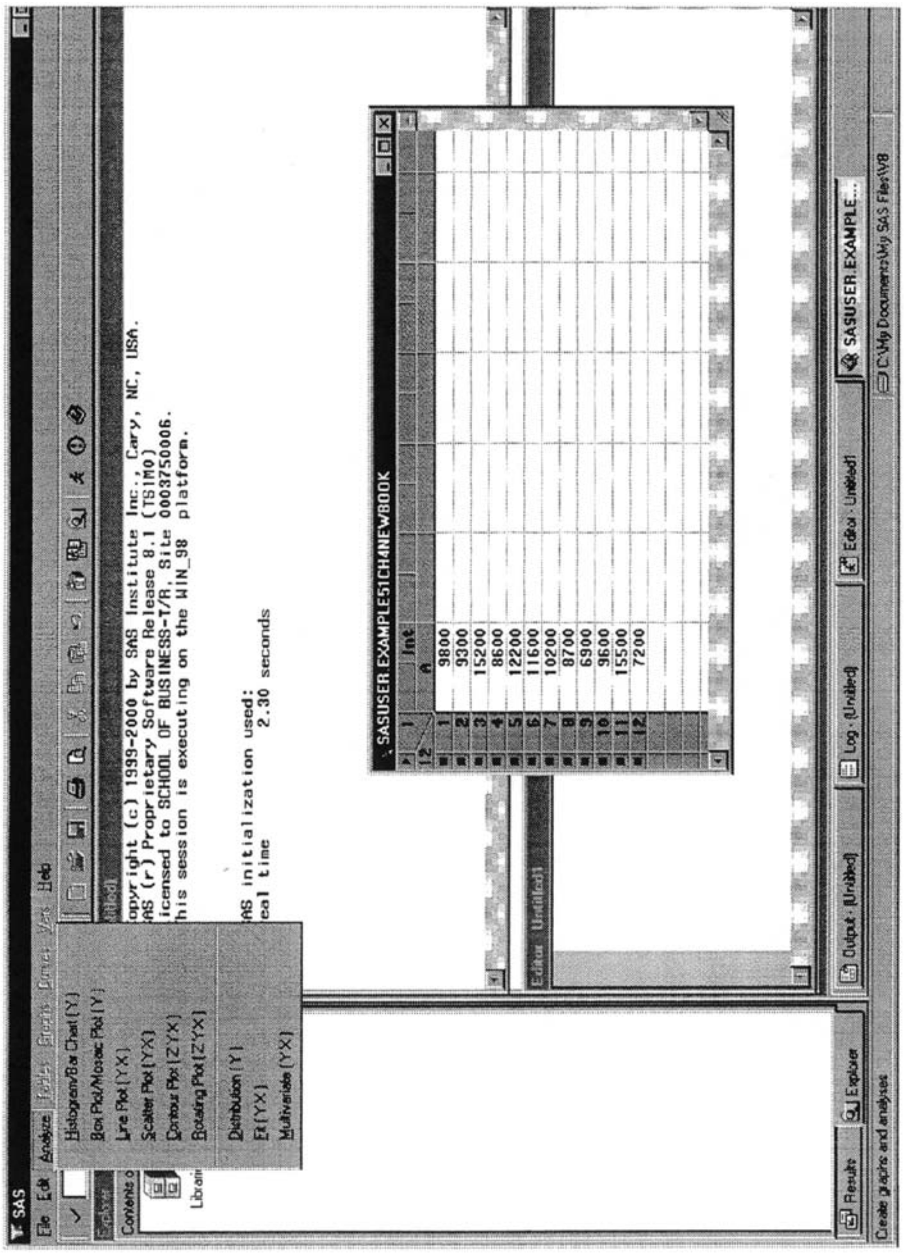


Fig. 5.1 SAS/Interactive Data Analysis goodness of fit for Example 5.1.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

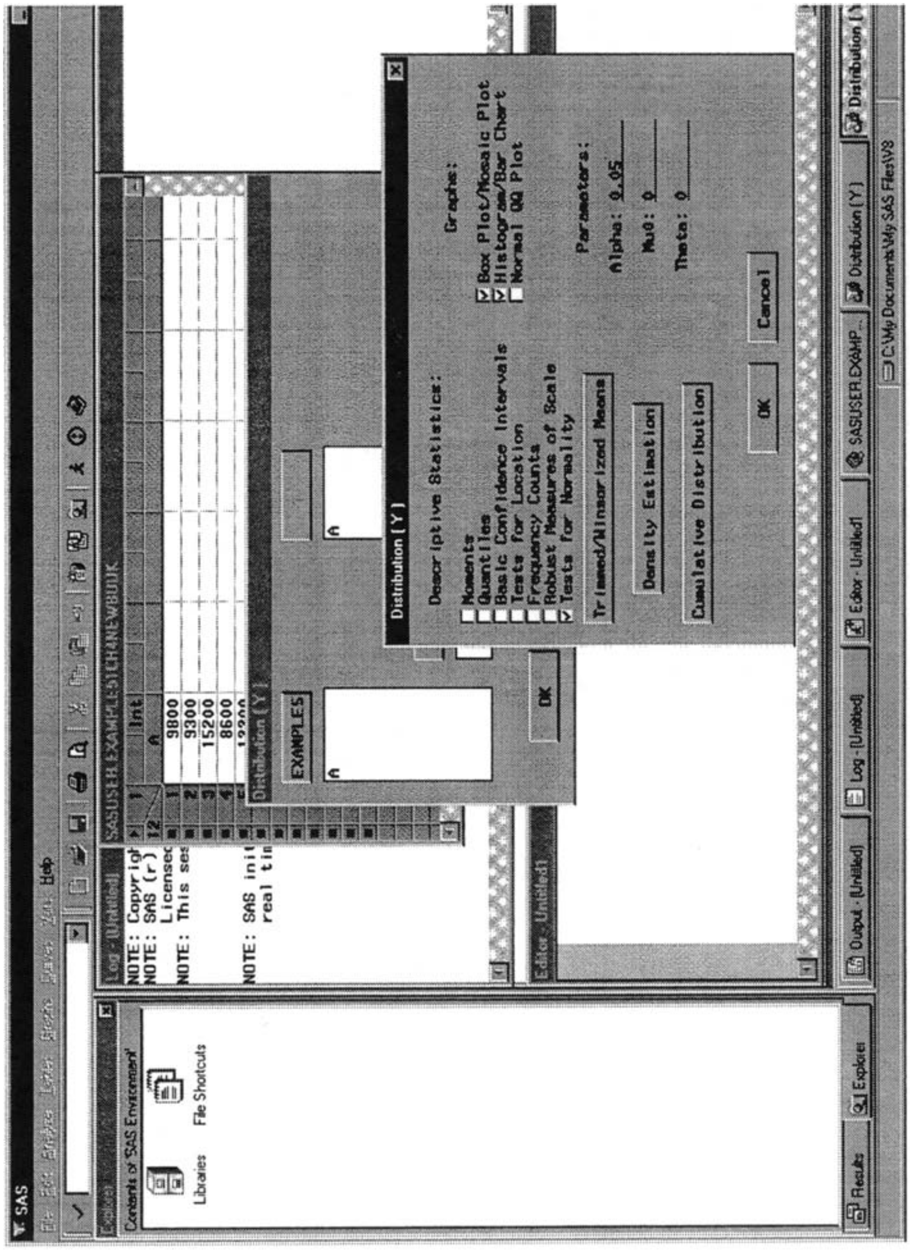


Fig. 5.2 SAS/Interactive Data Analysis goodness of fit for Example 5.1, continued.

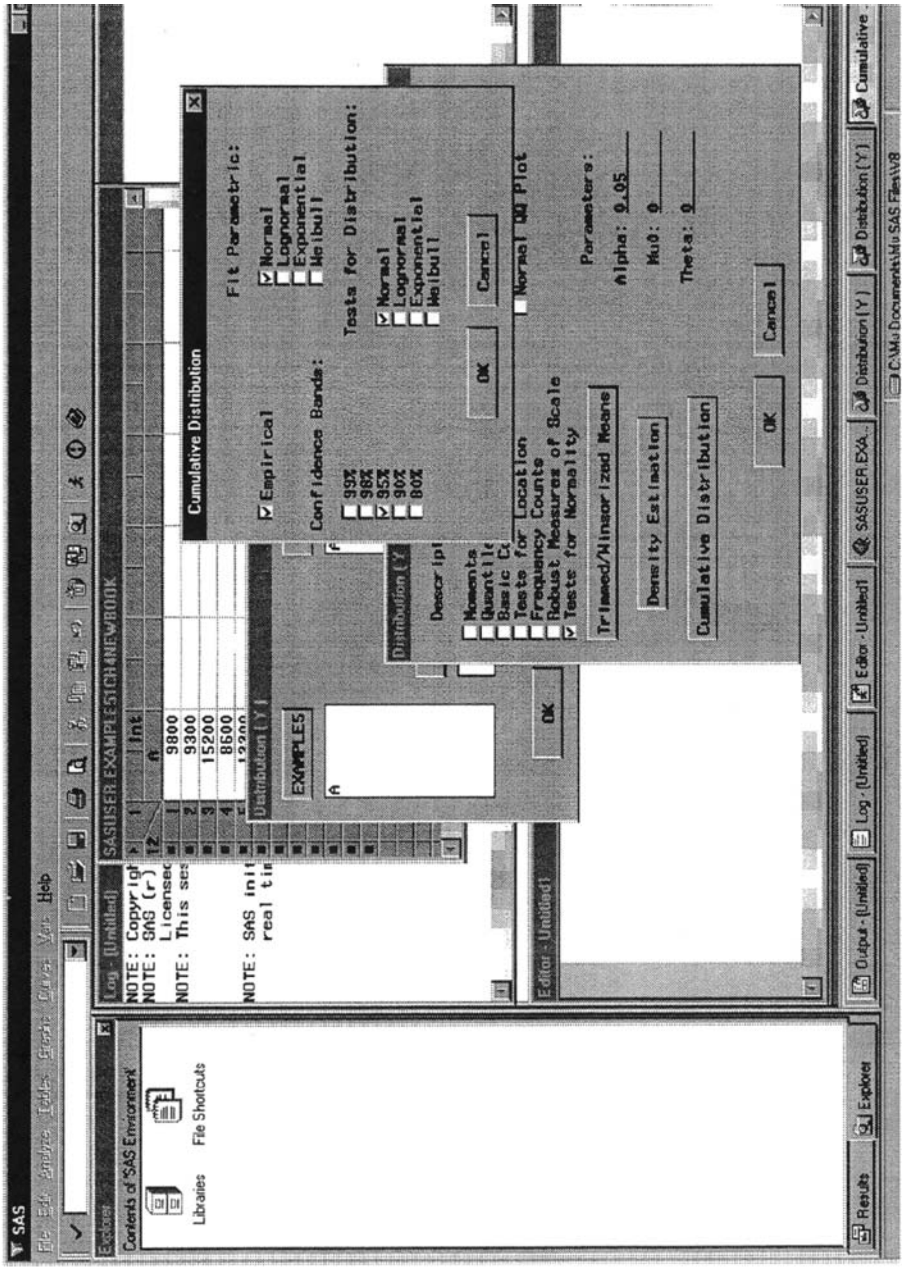


Fig. 5.3 SAS/Interactive Data Analysis of goodness of fit for Example 5.1, continued.

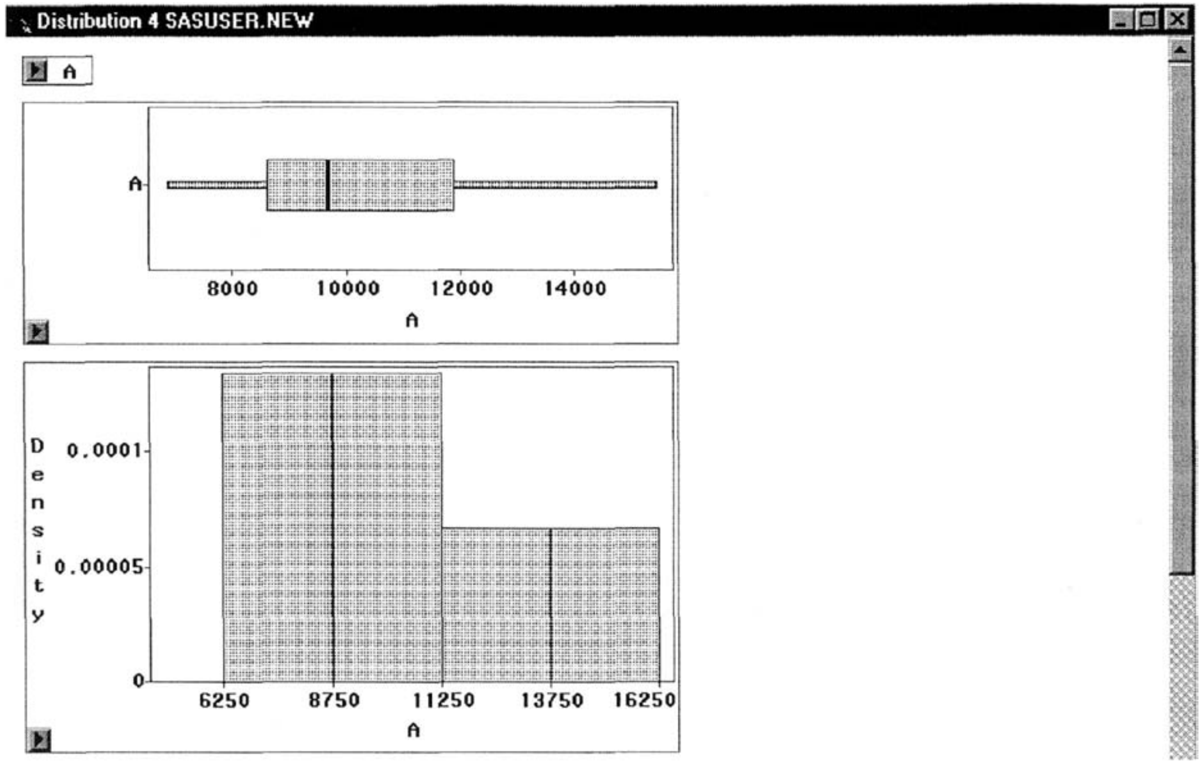


Fig. 5.4 SAS/Interactive Data Analysis of goodness of fit for Example 5.1, continued.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

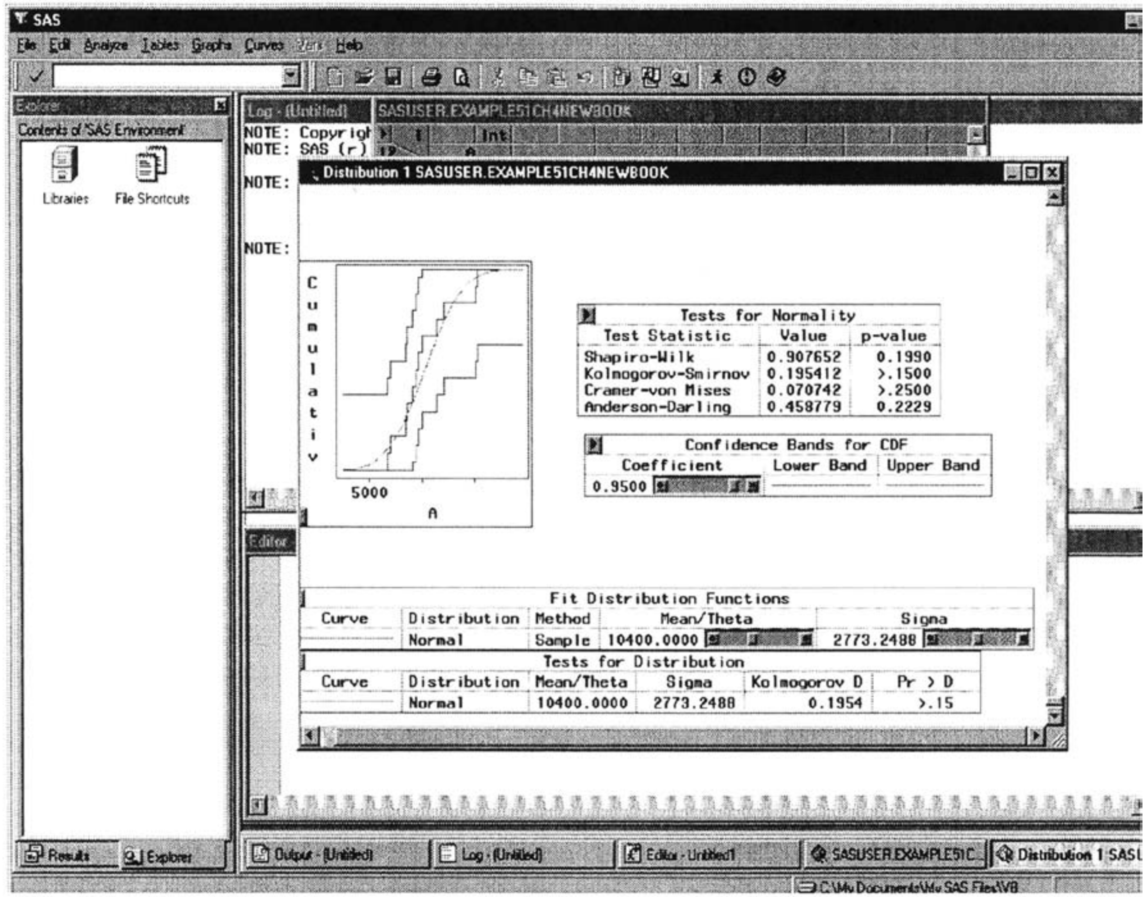


Fig. 5.5 SAS/Interactive Data Analysis of goodness of fit for Example 5.1, continued.

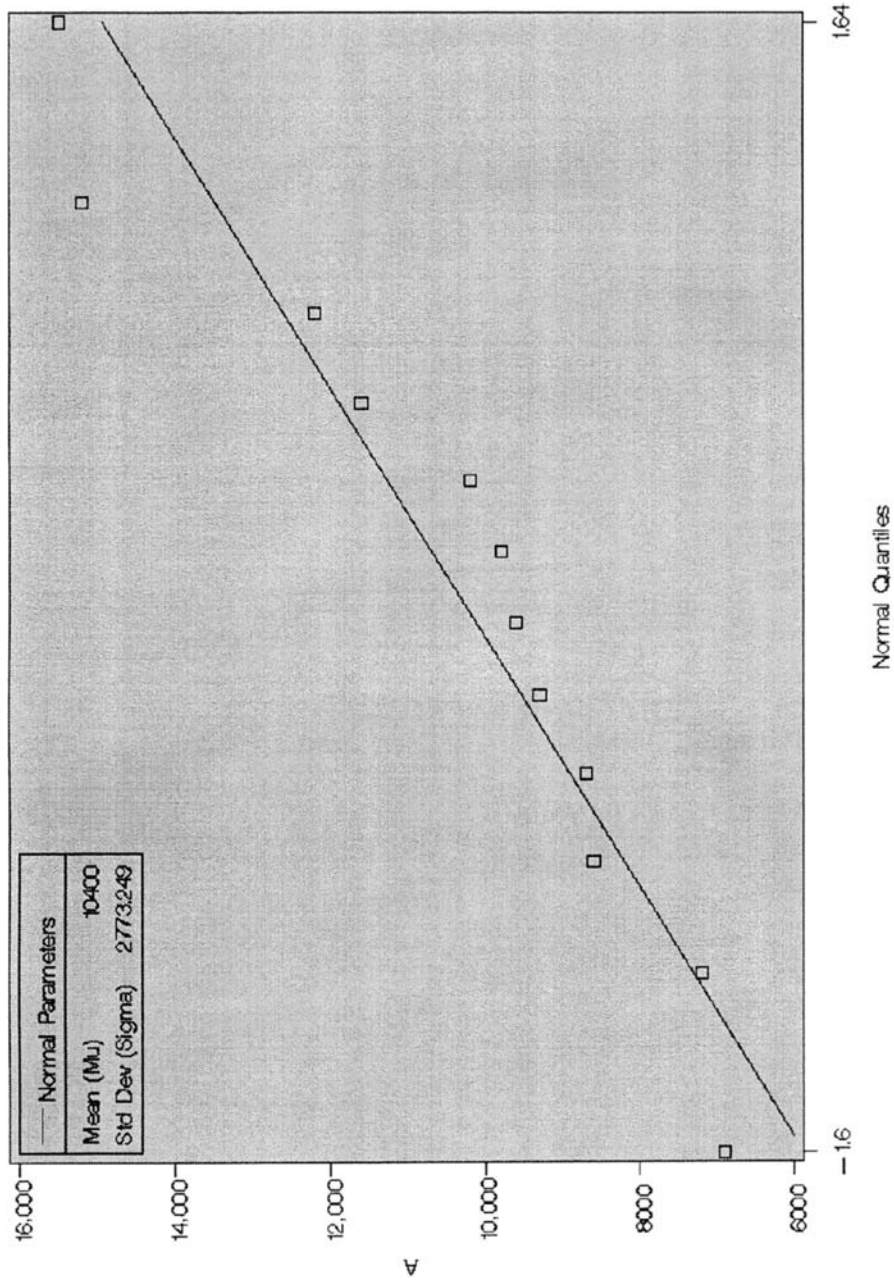


Fig. 5.6 SAS/Analyst Q-Q plot for Example 5.1.

the standardized variable $z_i = x_i/\bar{x}$, for each observed x_i, and calculates the usual K-S statistic between the empirical cdf and $\hat{F}_0(z_i)$. The appropriate rejection region is in the right tail and Table T in the Appendix gives the necessary critical values obtained by Monte Carlo simulations. This table is taken from Edgeman and Scott (1987) who used more samples to improve the accuracy of the original results given by Lilliefors (1969). Durbin (1975) provided exact quantiles.

Example 6.1 Test the null hypothesis that the data below arose from a one-parameter exponential distribution.

$$1.5, 2.3, 4.2, 7.1, 10.4, 8.4, 9.3, 6.5, 2.5, 4.6$$

Solution Since the mean of the exponential distribution is not specified under the null hypothesis, we estimate it from the data by $\bar{x} = 5.68$. The standardized variable is $z = x/5.68$. The calculations for the Lilliefors's test are summarized in Table 6.1. We find $D_n = 0.233$ and using Table T with $n = 10$, the approximate P value is greater than 0.10.

At the time of this writing, MINITAB does not provide any formal test for the exponential distribution, although some graphical procedures are available. The output from SAS/ANALYST for Example 6.1 is shown below. Note that the K-S test with the estimated mean is not referred to as Lilliefors's test but the numerical result agrees with ours. The approximate P value for Lilliefors's test (K-S test on the

Table 6.1 Calculations for Lilliefors's test in Example 6.1

X	z	$S_n(x)$	$\hat{F}_0(x) = F_0(z)$	$\lvert S_n(x) - F_0(z) \rvert$	$\lvert S_n(x-\varepsilon) - F_0(z) \rvert$
1.5	0.26	0.1000	0.2321	0.1321	0.2321
2.3	0.40	0.2000	0.3330	0.1330	0.2330
2.5	0.44	0.3000	0.3561	0.0561	0.1561
4.2	0.74	0.4000	0.5226	0.1226	0.2226
4.6	0.81	0.5000	0.5551	0.0551	0.1551
6.5	1.14	0.6000	0.6816	0.1816	0.1816
7.1	1.25	0.7000	0.7135	0.0135	0.1135
8.4	1.48	0.8000	0.7721	0.0279	0.0721
9.3	1.64	0.9000	0.8055	0.0945	0.0055
10.4	1.83	1.0000	0.8397	0.1603	0.0603

printout) is shown to be greater than 0.15. Thus, as with the hand calculations, we reach the same conclusion that there is not sufficient evidence to reject the null hypothesis of an exponential distribution. SAS uses internal tables that are similar to those given by D'Agostino and Stephens (1986) to calculate the P value. Linear interpolation is used in this table if necessary. SAS provides the values of two other test statistics, called the Cramér-von Mises and Anderson-Darling tests; each fails to reject the null hypothesis and the P values are about the same.

```
*************************************
 SAS/ANALYST SOLUTION FOR EXAMPLE 6.1
*************************************

          The UNIVARIATE Procedure
          Fitted Distribution for A

     Parameters for Exponential Distribution

         Parameter    Symbol    Estimate

         Threshold    Theta            0
         Scale        Sigma         5.68
         Mean                       5.68
         Std Dev                    5.68


   Goodness-of-Fit Tests for Exponential Distribution

Test                   ---Statistic----  ---p Value-----

Kolmogorov-Smirnov    D     0.23297622   Pr > D      >0.150
Cramer-von Mises      W-Sq  0.16302537   Pr > W-Sq    0.117
Anderson-Darling      A-Sq  0.94547938   Pr > A-Sq    0.120
```

We now redo Example 6.1 for the null hypothesis of the exponential distribution with mean specified as $\mu = 5.0$. This is a simple null hypothesis for which the original K-S test of Section 4.5 is applicable. The calculations are shown in Table 6.2 (p. 142). The K-S test statistic is $D_n = 0.2687$ with $n = 10$, and we do not reject the null hypothesis since Table F gives $P > 0.200$.

The SAS solution in this case is shown below. Each of the tests fails to reject the null hypothesis and the P values are about the same.

Table 6.2 Calculations for the K-S test with $\mu = 5.0$ for the data in Example 6.1

x	$z = x/\mu$	$S_n(x)$	$F_0(z)$	$\lvert S_n(x) - F_0(z) \rvert$	$\lvert S_n(x - \varepsilon) - F_0(z) \rvert$
1.5	0.30	0.1	0.2592	0.1592	0.2592
2.3	0.46	0.2	0.3687	0.1687	0.2687
2.5	0.50	0.3	0.3935	0.0935	0.1935
4.2	0.84	0.4	0.5683	0.1683	0.2683
4.6	0.92	0.5	0.6015	0.1015	0.2015
6.5	1.30	0.6	0.7275	0.1275	0.2275
7.1	1.42	0.7	0.7583	0.0583	0.1583
8.4	1.68	0.8	0.8136	0.0136	0.1136
9.3	1.86	0.9	0.8443	0.0557	0.0443
10.4	2.08	1.0	0.8751	0.1249	0.0249

```
***************************************************
  SAS/ANALYST SOLUTION FOR EXAMPLE 6.1 WITH MEAN 5
***************************************************

              The UNIVARIATE Procedure
              Fitted Distribution for A
        Parameters for Exponential Distribution

           Parameter    Symbol    Estimate

           Threshold    Theta            0
           Scale        Sigma            5
           Mean                          5
           Std Dev                       5


     Goodness-of-Fit Tests for Exponential Distribution

  Test                   ---Statistic----    -----p Value-----
  Kolmogorov-Smirnov     D     0.26871635    Pr > D      >0.250
  Cramer-von Mises       W-Sq  0.24402323    Pr > W-Sq    0.221
  Anderson-Darling       A-Sq  1.29014013    Pr > A-Sq    0.238
```

Finally suppose that we want to test the hypothesis that the population is exponential with mean $\mu = 3.0$. Again, this is a simple null hypothesis for which SAS provides three tests mentioned earlier and all of them reject the null hypothesis. However, note the difference in the magnitudes of the P values between the K-S test and the other two tests.

```
**************************************************
  SAS/ANALYST SOLUTION FOR EXAMPLE 6.1 WITH MEAN 3
**************************************************

                 The UNIVARIATE Procedure
                 Fitted Distribution for A

          Parameters for Exponential Distribution

               Parameter   Symbol   Estimate

               Threshold   Theta           0
               Scale       Sigma           3
               Mean                        3
               Std Dev                     3

      Goodness-of-Fit Tests for Exponential Distribution

Test                    ---Statistic----   -----p Value----
Kolmogorov-Smirnov      D     0.45340304   Pr > D      0.023
Cramer-von Mises        W-Sq  0.87408416   Pr > W-Sq   0.004
Anderson-Darling        A-Sq  4.77072898   Pr > A-Sq   0.004
```

Currently MINITAB does not provide a direct goodness-of-fit test for the exponential distribution but it does provide some options under a general visual approach. This is called probability plotting and is discussed in the next section.

4.7 VISUAL ANALYSIS OF GOODNESS OF FIT

With the advent of easily available computer technology, visual approaches to statistical data analysis have become popular. The subject is sometimes referred to as exploratory data analysis (EDA), championed by statisticians like John Tukey. In the context of goodness-of-fit tests, the EDA tools employed include dot plots, histograms, probability plots, and quantile plots. The idea is to use some graphics to gain a quick insight into the underlying distribution and then, if desired, carry out a follow-up analysis with a formal confirmatory test such as any of the tests covered earlier in this chapter. Dot plots and histograms are valuable exploratory tools and are discussed in almost all statistics books but the subject of probability and quantile plots is seldom covered, even though one of the key papers on the subject was published in the 1960s [Wilk and Gnanadesikan (1968)]. In this section we will present a brief discussion of these two topics. Fisher (1983) provided a good review of many graphical methods used in nonparametric statistics along with extensive references. Note that there are two-sample versions of each of these plots but we do not cover that topic here.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

In what follows we distinguish between the theoretical and the empirical versions of a plot. The theoretical version is presented to understand the idea but the empirical version is the one that is implemented in practice. When there is no chance of confusion, the empirical plot is referred to as simply the plot.

Two types of plots are popular in practice. The first is the so-called probability plot, which is actually a probability versus probability plot, or a P-P plot. This plot is also called a percent-percent plot, for obvious reasons. In general terms, the theoretical P-P plot is the graph of a cdf $F(x)$ versus a cdf $G(x)$ for all values of x. Since the cdf's are probabilities, the P-P plot is conveniently confined to the unit square. If the two cdfs are identical, the theoretical P-P plot will be the main diagonal, the 45 degree line through the origin.

The second type of plot is the so-called quantile plot, which is actually a quantile versus quantile plot, or a Q-Q plot. The theoretical Q-Q plot is the graph of the quantiles of a cdf F versus the corresponding quantiles of a cdf G, that is, the graph $[F^{-1}(p), G^{-1}(p)]$ for $0 < p < 1$. If the two cdf's are identical, the theoretical Q-Q plot will be the main diagonal, the 45-degree line through the origin. If $F(x) = G(\frac{x-\mu}{\sigma})$, it is easily seen that $F^{-1}(p) = \mu + \sigma G^{-1}(p)$, so that the pth quantiles of F and G have a linear relationship. Thus, if two distributions differ only in location and/or scale, the theoretical Q-Q plot will be a straight line with slope σ and intercept μ.

In a goodness-of-fit problem, there is usually a specified target cdf, say F_0. Then the theoretical Q-Q plot is the plot $[F_0^{-1}(p), F_X^{-1}(p)]$, $0 < p < 1$. Since F_X is unknown, we can estimate it with the empirical cdf based on a random sample of size n, say S_n. Noting that the function S_n jumps only at the ordered values $X_{(i)}$, the empirical Q-Q plot is simply the plot of $F_0^{-1}(i/n)$ on the horizontal axis versus $S_n^{-1}(i/n) = X_{(i)}$ on the vertical axis, for $i = 1, 2, \ldots, n$. As noted before, F_0 is usually taken to be the standardized form of the hypothesized cdf, so that to establish the Q-Q plot (location and/or scale), underlying parameters do not need to be specified. This is one advantage of the Q-Q plot. The quantities $a_i = i/n$ are called plotting positions. At $i = n$, there is a problem since $a_n = F_0^{-1}(1) = \infty$; modified plotting positions have been considered, with various objectives. One simple choice is $a_i = i - 0.5/n$; other choices include $a_i = i/n + 1$ and $a_i = (i - 0.375)/n + 0.25$, the latter being highly recommended by Blom (1958). We found that many statistical software package graph $\left[F_0^{-1}((i - 0.375)/(n + 0.25)),\ X_{(i)}\right]$ as the empirical Q-Q plot. For a given standardized cdf F_0, the goodness-of-fit null hypothesis $F_X = F_0$ is not rejected if this plot is approximately a straight line through

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

the origin. Departures from this line suggest the types of differences that could exist between F_X and F_0. For example, if the plot resembles a straight line but with a nonzero intercept or with a slope other than 45 degrees, a location-scale model is indicated. This means F_X belongs to the specified family of distributions but the location and the scale parameters of F_X, namely μ and σ, are different from the standard values. When the empirical Q-Q plot is reasonably linear, the slope and the intercept of the plot can be used to estimate the scale and location parameter, respectively. When F_0 is taken to be the standard normal distribution, the Q-Q plot is called a normal probability plot. When F_0 is taken to be the standard exponential distribution (mean $= 1$), the Q-Q plot is called an exponential probability plot.

In summary, either the empirical P-P or Q-Q plot can be used as an informal tool for the goodness-of-fit problem but the Q-Q plot is more popular. If the plots appear to be close to the 45 degree straight line through the origin, the null hypothesis $F_X = F_0$ is tentatively accepted. If the Q-Q plot is close to some other straight line, then F_X is likely to be in the hypothesized location-scale family (as F_0) and the unknown parameters can be estimated from the plot. For example, if a straight line is fitted to the empirical Q-Q plot, the slope and the intercept of the line would estimate the unknown scale and the location parameter, respectively; then the estimated distribution is $\widehat{F}_X = F_0\left(\frac{x-\text{intercept}}{\text{slope}}\right)$. An advantage of the Q-Q plot is that the underlying parameters do not need to be specified since F_0 is usually taken to be the standard distribution in a family of distributions. By contrast, the construction of a P-P plot requires specification of the underlying parameters, so that the theoretical cdf can be evaluated at the ordered data values. The P-P plot is more sensitive to the differences in the middle part of the two distributions (the data distribution and the hypothesized distribution), whereas the Q-Q plot is more sensitive to the differences in the tails of the two distributions.

One potential issue with using plots in goodness-of-fit problems is that the interpretation of a plot, with respect to linearity or near linearity, is bound to be somewhat subjective. Usually a lot of experience is necessary to make the judgment with a reasonable degree of confidence. To make such an assessment more objective, several proposals have been made. One is based on the "correlation coefficient" between the x and y coordinates; see Ryan and Joiner (1976) for a test in the context of a normal probability plot. For more details, see D'Agostino and Stephens (1986, Chap. 5).

Table 7.1 Calculations for normal and exponential Q-Q plot for data in Example 6.1

Ordered data y	i	*Plotpos* $a_i = \frac{i-.375}{10.25}$	*Standard normal quantiles* $\Phi^{-1}(a_i)$	*Standard exponential quantiles* $-\ln(1-a_i)$
1.5	1	0.060976	− 1.54664	0.062914
2.3	2	0.158537	− 1.00049	0.172613
2.5	3	0.256098	− 0.65542	0.295845
4.2	4	0.353659	− 0.37546	0.436427
4.6	5	0.451220	− 0.12258	0.600057
4.5	6	0.548780	0.12258	0.795801
7.1	7	0.646341	0.37546	1.039423
8.4	8	0.743902	0.65542	1.362197
9.3	9	0.841463	1.00049	1.841770
10.4	10	0.939024	1.54664	2.797281

Example 7.1 For the sample data given in Example 6.1 using $a_i = (i - 0.375)/(n + 0.25)$, the calculations for a normal and exponential Q-Q plots are shown in Table 7.1. The two Q-Q plots are plotted in EXCEL and are shown in Figures 7.1 and 7.2. In each case a least-squares line is fitted to the plot. The slope and the intercept of

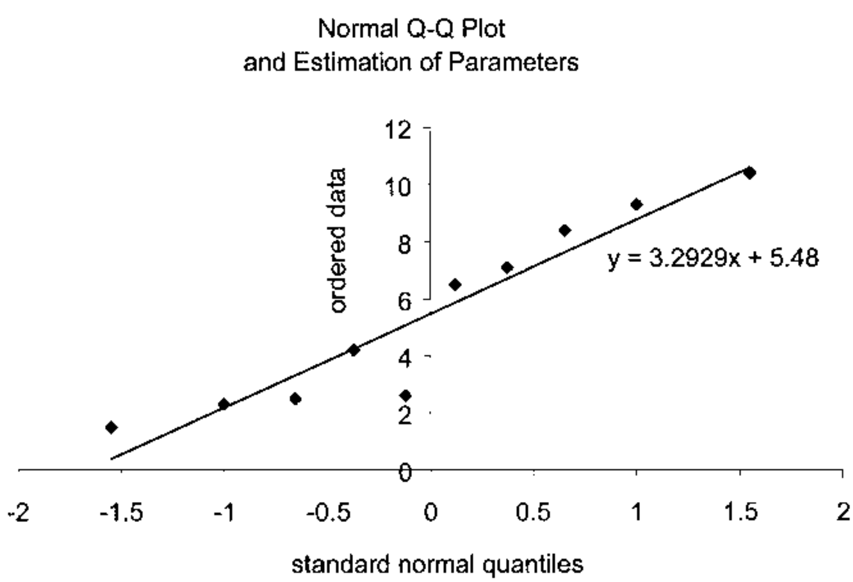


Fig. 7.1 Normal Q-Q plot for Example 7.1.

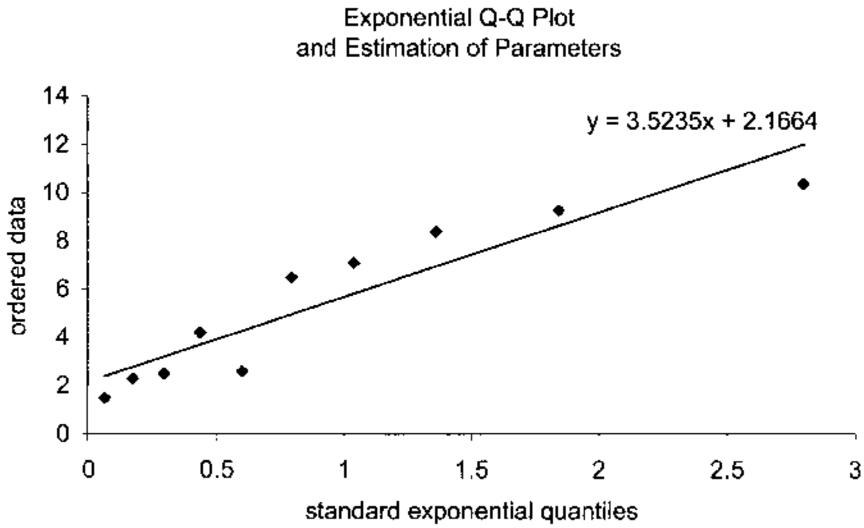


Fig. 7.2 Exponential Q-Q plot for Example 7.1.

the scale and the location parameter under the respective model. For these data it appears that the normal distribution with mean 5.48 and standard deviation 3.3 provides a better fit than the exponential distribution with mean 3.52.

4.8 SUMMARY

In this chapter we presented procedures designed to help identify the population from which a random sample is drawn. The primary test criteria are the normalized sum of squares of deviations, and the difference of the cumulative distribution functions, between the hypothesized and observed (sample) distributions. Chi-square tests are based on the first criterion and the K-S type tests are based on the second (including the Lilliefors's tests).

The chi-square test is specifically designed for use with categorical data, while the K-S statistics are for random samples from continuous populations. However, when the data are not categorical as collected, these two goodness-of-fit tests can be used interchangeably. The reader is referred to Goodman (1954), Birnbaum (1952), and Massey (1951c) for discussions of their relative merits. Only a brief comparison will be made here, which is relevant whenever raw ungrouped measurement data are available.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The basic difference between the two tests is that chi square is sensitive to vertical deviations between the observed and expected histograms, whereas K-S type procedures are based on vertical deviations between the observed and expected cumulative distribution functions. However, both types of deviations are useful in determining goodness-of-fit and probably are equally informative. Another obvious difference is that chi square requires grouped data whereas K-S does not. Therefore, when the hypothesized distribution is continuous, the K-S test allows us to examine the goodness-of-fit for each of the n observations, instead of only for the k classes, where $k \leqslant n$. In this sense, the K-S type procedures make more complete use of the available data. Further, the chi-square statistic is affected by the number of classes and their widths, which are often chosen arbitrarily by the experimenter.

One of the primary advantages of the K-S test is that the exact sampling distribution of D_n is known or can be simulated and tabulated when all parameters are specified, whereas the sampling distribution of Q is only approximately chi square for any finite n. The K-S test can be applied for any size sample, while the chi-square test should be used only for n large and each expected cell frequency not too small. When cells must be combined for chi-square application, the calculated value of Q is no longer unique, as it is affected by the scheme of combination. The K-S statistic is much more flexible than chi square, since it can be used in estimation to find minimum sample size and confidence bands. With the one-sided D_n^+ or D_n^- statistics, we can test for deviations in a particular direction, whereas the chi-square is always concerned equally with differences in either direction. In most cases, K-S is easier to apply.

However, the chi-square test also has some advantages over K-S. A hypothesized distribution which is discrete presents no problems for the chi-square test, while the exact properties of D_n are violated by the lack of continuity. As already stated, however, this is a minor problem with D_n which can generally be eliminated by replacing equalities by inequalities in the probabilities. Perhaps the main advantage of the chi-square test is that by simply reducing the number of degrees of freedom and replacing unknown parameters by consistent estimators, a goodness-of-fit test can be performed in the usual manner even when the hypothesized distribution is not completely specified. If the hypothesized $F_0(x)$ in D_n contains unspecified parameters which are estimated from the data, we obtain an estimate $\hat{D}_n$ whose sampling distribution is different from that of D_n. The test is conservative when the D_n critical values are used.

As for relative performance, the power functions of the two statistics depend on different quantities. If $F_0(x)$ is the hypothesized cumulative distribution and $F_X(x)$ the true distribution, the power of K-S depends on

$$\sup_x |F_X(x) - F_0(x)|$$

while the power of chi square depends on

$$\sum_{i=0}^{k} \frac{\{[F_X(a_{i+1}) - F_X(a_i)] - [F_0(a_{i+1}) - F_0(a_i)]\}^2}{F_0(a_{i+1}) - F_0(a_i)}$$

where the a_i are the class limits in the numerical categories.

The power of the chi-square test can be improved by clever grouping in some situations. In particular, Cochran (1952) and others have shown that a choice of intervals which provide equal expected frequencies for all classes is a good procedure in this respect besides simplifying the computations. The number of classes k can be chosen such that the power is maximized in the vicinity of the point where power equals 0.5. This procedure also eliminates the arbitrariness of grouping. The expression for Q in (2.1) reduces to $(k \sum F_i^2 - n^2)/n$ when $e_i = n/k$ for $i = 1, 2, \ldots, k$.

Many studies of power comparisons have been reported in the literature over the years. Kac, Kiefer, and Wolfowitz (1955) showed that the K-S test is asymptotically more powerful than the chi-square test when testing for a completely specified normal distribution. Further, when the sample size is small, the K-S provides an exact test while the chi-square does not.

When the hypothesized distribution is the normal or exponential and the parameters are specified (the null hypothesis is simple), the K-S test based on Table F gives an exact goodness-of-fit test. This test is conservative when parameters need to be estimated (the null hypothesis is composite). In these cases, the modified version of the D_n statistic sometimes known as the Lilliefors's test statistic should be used. The exact mathematical-statistical derivation of the distribution of this test statistic is often very complicated but Monte Carlo estimates of the percentiles of the null distribution can be

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

obtained. The tables given in Lilliefors (1967, 1969) were generated in this manner, as are our Tables O and T. Edgeman and Scott (1987) gave a step-by-step algorithm that included goodness-of-fit testing for the lognormal, the Rayleigh, Weibull and the two-parameter exponential distribution.

Iman (1982) provided graphs for performing goodness-of-fit tests for the normal and the exponential distributions with unspecified parameters. These graphs are in the form of confidence bands based on Lilliefors's (1967) critical values for the normal distribution test and Durbin's (1975) critical values for the exponential distribution test. On a graph with these bands, the empirical distribution function of the standardized variable $S_n(z)$ is plotted on the vertical axis as a step function in order to carry out the test. If this plot lies entirely within the confidence bands, the null hypothesis is not rejected. These graphs can be obtained by running some MINITAB macro programs.

Finally, we have the option of using P-P or Q-Q plots to glean information about the population distribution. The interpretation is usually subjective, however.

PROBLEMS

4.1. Two types of corn (golden and green-striped) carry recessive genes. When these were crossed, a first generation was obtained which was consistently normal (neither golden nor green-striped). When this generation was allowed to self-fertilize, four distinct types of plants were produced: normal, golden, green-striped, and golden-green-striped. In 1200 plants this process produced the following distribution:

Normal: 670
Golden: 230
Green-striped: 238
Golden-green-striped: 62

A monk named Mendel wrote an article theorizing that in a second generation of such hybrids, the distribution of plant types should be in a 9:3:3:1 ratio. Are the above data consistent with the good monk's theory?

4.2. A group of four coins is tossed 160 times, and the following data are obtained:

Number of heads	0	1	2	3	4
Frequency	16	48	55	33	8

Do you think the four coins are balanced?

4.3. A certain genetic model suggests that the probabilities for a particular trinomial distribution are, respectively, $\theta_1 = p^2, \theta_2 = 2p(1-p)$, and $\theta_3 = (1-p)^2, 0 < p < 1$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Assume that X_1, X_2, and X_3 represent the respective frequencies in a sample of n independent trials and that these numbers are known. Derive a chi-square goodness-of-fit test for this trinomial distribution if p is unknown.

4.4. According to a genetic model, the proportions of individuals having the four blood types should be related by

Type 0: q^2
Type A: $p^2 + 2pq$
Type B: $r^2 + 2qr$
Type AB: $2pr$

where $p + q + r = 1$. Given the blood types of 1000 individuals, how would you test the adequacy of the model?

4.5. If individuals are classified according to gender and color blindness, it is hypothesized that the distribution should be as follows:

	Male	*Female*
Normal	$p/2$	$p^2/2 + pq$
Color blind	$q/2$	$q^2/2$

for some $p + q = 1$, where p denotes the proportion of defective genes in the relevant population and therefore changes for each problem. How would the chi-square test be used to test the adequacy of the general model?

4.6. Show that in general, for Q defined as in (2.1),

$$E(Q) = E\left[\sum_{i=1}^{k} \frac{(F_i - e_i)^2}{e_i}\right] = \sum_{i=1}^{k}\left[\frac{n\theta_i(1-\theta_i)}{e_i} + \frac{(n\theta_i - e_i)^2}{e_i}\right]$$

From this we see that if the null hypothesis is true, $n\theta_i = e_i$ and $E(Q) = k - 1$, the mean of the chi-square distribution.

4.7. Show algebraically that where $e_i = n\theta_i$ and $k = 2$, we have

$$Q = \sum_{i=1}^{2} \frac{(F_i - e_i)^2}{e_i} = \frac{(F_1 - n\theta_1)^2}{n\theta_1(1-\theta_1)}$$

so that when $k = 2$, $\sqrt{Q}$ is the statistic commonly used for testing a hypothesis concerning the parameter of the binomial distribution for large samples. By the central-limit theorem, $\sqrt{Q}$ approaches the standard normal distribution as $n \to \infty$ and the square of any standard normal variable is chi-square-distributed with 1 degree of freedom. Thus we have an entirely different argument for the distribution of Q when $k = 2$.

4.8. Give a simple proof that D_n, D_n^+, and D_n^- are completely distribution-free for any continuous F_X by appealing to the transformation $u = F_X(x)$ in the initial definitions of D_n, D_n^+, and D_n^-.

4.9. Prove that

$$D_n^- = \max\left\{\max_{1 \leqslant i \leqslant n}\left[F_X(X_{(i)}) - \frac{i-1}{n}\right], 0\right\}$$

4.10. Prove that the probability distribution of D_n^- is identical to the distribution of D_n^+:

(*a*) Using a derivation analogous to Theorem 3.4
(*b*) Using a symmetry argument

4.11. Using Theorem 3.3, verify that

$$\lim_{n\to\infty} P\big(D_n > 1.07/\sqrt{n}\big) = 0.20$$

4.12. Find the minimum sample size n required such that $P(D_n < 0.05) \geqslant 0.99$.

4.13. Use Theorem 3.4 to verify directly that $P(D_5^+ > 0.447) = 0.10$. Calculate this same probability using the expression given in (3.5).

4.14. Related goodness-of-fit test. The Cramér-von Mises type of statistic is defined for continuous $F_X(x)$ by

$$\omega_n^2 = \int_{-\infty}^{\infty} [S_n(x) - F_X(x)]^2 f_X(x)\,dx$$

(*a*) Prove that ω_n^2 is distribution free.
(*b*) Explain how ω_n^2 might be used for a goodness-of-fit test.
(*c*) Show that

$$n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F_X\left(X_{(i)} - \frac{2i-1}{n}\right)\right]^2$$

This statistic is discussed in Cramér (1928), von Mises (1931), Smirnov (1936), and Darling (1957).

4.15. Suppose we want to estimate the cumulative distribution function of a continuous population using the empirical distribution function such that the probability is 0.90 that the error of the estimate does not exceed 0.25 anywhere. How large a sample size is needed?

4.16. If we wish to estimate a cumulative distribution within 0.20 units with probability 0.95, how large should n be?

4.17. A random sample of size 13 is drawn from an unknown continuous population $F_X(x)$, with the following results after array:

$$3.5, 4.1, 4.8, 5.0, 6.3, 7.1, 7.2, 7.8, 8.1, 8.4, 8.6, 9.0$$

A 90% confidence band is desired for $F_X(x)$. Plot a graph of the empirical distribution function $S_n(x)$ and resulting confidence bands.

4.18. In a vibration study, a random sample of 15 airplane components were subjected to severe vibrations until they showed structural failures. The data given are failure times in minutes. Test the null hypothesis that these observations can be regarded as a sample from the exponential population with density function $f(x) = e^{-x/10}/10$ for $x \geqslant 0$.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

1.6, 10.3, 3.5, 13.5, 18.4, 7.7, 24.3, 10.7, 8.4, 4.9, 7.9, 12.0, 16.2, 6.8, 14.7

4.19. For the data given in Example 6.1 use the most appropriate test to see if the distribution can be assumed to be normal with mean 10,000 and standard deviation 2,000.

4.20. The data below represent earnings per share (in dollars) for a random sample of five common stocks listed on the New York Stock Exchange.

1.68, 3.35, 2.50, 6.23, 3.24

(*a*) Use the most appropriate test to see if these data can be regarded as a random sample from a normal distribution.
(*b*) Use the most appropriate test to see if these data can be regarded as a random sample from a normal distribution with $\mu = 3, \sigma = 1$.
(*c*) Determine the sample size required to use the empirical distribution function to estimate the unknown cumulative distribution function with 95% confidence such that the error in the estimate is (i) less than 0.25, (ii) less than 0.20.

4.21. It is claimed that the number of errors made by a typesetter is Poisson distributed with an average rate of 4 per 1000 words set. One hundred random samples of sets of 1000 words from this typesetter output are examined and the numbers of errors are counted as shown below. Are these data consistent with the claim?

No. of errors	0	1	2	3	4	5
No. of samples	10	16	20	28	12	14

4.22. For the original data in Example 3.1 (not the square roots), test the null hypothesis that they come from the continuous uniform distribution, using level 0.01.

4.23. Use the D_n statistic to test the null hypothesis that the data in Example 2.1:

(*a*) Come from the Poisson distribution with $\mu = 1.5$
(*b*) Come from the binomial distribution with $n = 13,\ p = 0.1$

These tests will be conservative because both hypothesized distributions are discrete.

4.24. Each student in a class of 18 is asked to list three people he likes and three he dislikes and label the people 0, 1, 2, 3, 4, 5 according to how much he likes them, with 0 denoting least liked and 5 denoting most liked. From this list each student selects the number assigned to the person he thinks is the wealthiest of the six. The results in the form of an array are as follows:

0, 0, 0, 0, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 5

Test the null hypothesis that the students are equally likely to select any of the numbers 0, 1, 2, 3, 4, 5, using the most appropriate test and the 0.05 level of significance.

4.25. During a 50-week period, demand for a certain kind of replacement part for TV sets was distributed as shown below. Find the theoretical distribution of weekly demands for a Poisson model with the same mean as the given data and perform an appropriate goodness-of-fit test.

Weekly demand	*Number of weeks*
0	28
1	15
2	6
3	1
More than 3	0
	50

4.26. Suppose that monthly collections for home delivery of the *New York Times* in a large suburb of New York are approximately normally distributed with mean \$150 and standard deviation \$20. A random sample of 10 delivery persons in a nearby suburb is taken; the arrayed data for monthly collections in dollars are:

$$90, 106, 109, 117, 130, 145, 156, 170, 174, 190$$

Test the null hypothesis that the same normal distribution model applies to this suburb, using the most appropriate test.

4.27. A bank frequently makes large installment loans to builders. At any point in time, outstanding loans are classified in the following four repayment categories:

A: Current
B: Up to 30 days delinquent
C: 30–60 days delinquent
D: Over 60 days delinquent

The bank has established the internal standard that these loans are "in control" as long as the percentage in each category is as follows:

$$A{:}\,80\% \qquad B{:}\,12\% \qquad C{:}\,7\% \qquad D{:}\,1\%$$

They make frequent spot checks by drawing a random sample of loan files, noting their repayment status at that time and comparing the observed distribution with the standard for control. Suppose a sample of 500 files produces the following data on number of loans in each repayment category:

$$A{:}\,358 \qquad B{:}\,83 \qquad C{:}\,44 \qquad D{:}\,15$$

Does it appear that installment loan operations are under control at this time?

4.28. Durtco Incorporated designs and manufactures gears for heavy-duty construction equipment. One such gear, 9973, has the following specifications:

(*a*) Mean diameter 3.0 in.
(*b*) Standard deviation 0.001 in.
(*c*) Output normally distributed

The production control manager has selected a random sample of 500 gears from the inventory and measured the diameter of each. Nothing more has been done to the data. How would you determine statistically whether gear 9973 meets the specifications?

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Be brief but specific about which statistical procedure to use and why it is preferred and outline the steps in the procedure.

4.29. Compare and contrast the chi-square and Kolmogorov-Smirnov goodness-of-fit procedures.

4.30. For the data x: $1.0, 2.3, 4.2, 7.1, 10.4$, use the most appropriate procedure to test the null hypothesis that the distribution is

(*a*) Exponential $F_0(x) = 1 - e^{-\lambda x}$ (estimate λ by $1/\bar{x}$)
(*b*) Normal

In each part, carry the parameter estimates to the nearest hundredth and the distribution estimates to the nearest ten thousandth.

4.31. A statistics professor claims that the distribution of final grades from A to F in a particular course invariably is in the ratio 1:3:4:1:1. The final grades this year are 26 A's, 50 B's, 80 C's, 35 D's, and 10 F's. Do these results refute the professor's claim?

4.32. The design department has proposed three different package designs for the company's product; the marketing manager claims that the first design will be twice as popular as the second design and that the second design will be three times as popular as the third design. In a market test with 213 persons, 111 preferred the first design, 62 preferred the second design, and the remainder preferred the third design. Are these results consistent with the marketing manager's claim?

4.33. A quality control engineer has taken 50 samples, each of size 13, from a production process. The numbers of defectives are recorded below.

Number of defects	*Sample frequency*
0	9
1	26
2	9
3	4
4	1
5	1
6 or more	0

(*a*) Test the null hypothesis that the number of defectives follows a Poisson distribution.
(*b*) Test the null hypothesis that the number of defectives follows a binomial distribution.
(*c*) Comment on your answers in (*a*) and (*b*).

4.34. Ten students take a test and their scores (out of 100) are as follows:

$95, 80, 40, 52, 60, 80, 82, 58, 65, 50$

Test the null hypothesis that the cumulative distribution function of the proportion of right answers a student gets on the test is

$$F_0(x) = \begin{cases} 0 & x < 1 \\ x^2(3 - 2x) & 0 \leqslant x \leqslant 1 \\ 1 & x > 1 \end{cases}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

5
One-Sample and Paired-Sample Procedures

5.1 INTRODUCTION

In the general one-sample problem, the available data consist of a single set of observations, usually a random sample, from a cdf F_X on which inferences can be based regarding some aspect of F_X. The tests for randomness in Chapter 3 relate to inferences about a property of the joint probability distribution of a set of sample observations which are identically distributed but possibly dependent, i.e., the probability distribution of the data. The hypothesis in a goodness-of-fit study in Chapter 4 is concerned with the univariate population distribution from which a set of independent variables is drawn. These hypotheses are so general that no analogous counterparts exist within the realm of parametric statistics. Thus these problems are more suitable to be viewed under nonparametric procedures. In a classical one-sample inference problem, the single-sample data are used to obtain information about some particular aspect of the population distribution, usually one or more of its parameters. Nonparametric techniques are useful here too, particularly when a location parameter is of interest.

In this chapter we shall be concerned with the nonparametric analog of the normal-theory test (variance known) or Student's t test (variance unknown) for the hypotheses $H_0: \mu = \mu_0$ and $H_0: \mu_X - \mu_Y = \mu_D = \mu_0$ for the one-sample and paired-sample problems, respectively. The classical tests are derived under the assumption that the single population or the population of differences of pairs is normal. For the nonparametric tests, however, only certain continuity assumptions about the populations need to be postulated to determine sampling distributions of the test statistics. The hypotheses here are concerned with the median or some other quantile rather than the mean as the location parameter, but both the mean and the median are good indexes of central tendency and they do coincide for symmetric populations. In any population, the median always exists (which is not true for the mean) and it is more robust as an estimate of location. The procedures covered here include confidence intervals and tests of hypotheses about any specified quantile. The case of the median is treated separately and the popular sign test and the Wilcoxon signed-rank test, including both hypothesis testing and confidence interval techniques, are presented. The complete discussion in each case will be given only for the single-sample case, since with paired-sample data once the differences of observations are formed, we have essentially only a single sample drawn from the population of differences and thus the methods of analysis are identical.

We also introduce rank-order statistics and present a measure of the relationship between ranks and variate values.

5.2 CONFIDENCE INTERVAL FOR A POPULATION QUANTILE

Recall from Chapter 2 that a quantile of a continuous random variable X is a real number that divides the area under the probability density function into two parts of specified amounts. Only the area to the left of the number need be specified since the entire area is equal to 1. Let F_X be the underlying cdf and let κ_p, for all $0 < p < 1$, denote the pth quantile, or the 100pth percentile, or the quantile of order p of F_X. Thus, κ_p is defined to be any real number which is a solution to the equation $F_X(\kappa_p) = p$, and in terms of the *quantile function*, $\kappa_p = Q_X(p) = F_X^{-1}(p)$. We shall assume here that a unique solution (inverse) exists, as would be the case for a strictly increasing function F_X. Note that κ_p is a parameter of the population F_X, and to emphasize

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

this point we use the Greek letter κ_p instead of the Latin letter $Q_X(p)$ used before in Chapter 2. For example, $\kappa_{0.50}$ is the median of the distribution, a measure of central tendency.

First we consider the problem where a confidence interval estimate of the parameter κ_p is desired for some specified value of p, given a random sample $X_1, X_2, \ldots, X_n$ from the cdf F_X. As discussed in Chapter 2, a natural point estimate of κ_p would be the pth sample quantile, which is the (np)th-order statistic, provided of course that np is an integer. For example, since $100p$ percent of the population values are less than or equal to the pth population quantile, the estimate of κ_p is that value from a random sample such that $100p$ percent of the sample values are less than or equal to it. We define $X_{(r)}$ to be the pth sample quantile where r is defined by

$$r = \begin{cases} np & \text{if } np \text{ is an integer} \\ [np + 1] & \text{if } np \text{ is not an integer} \end{cases}$$

and [x] denotes the largest integer not exceeding x. This is just a convention adopted so that we can handle situations where np is not an integer. Other conventions are sometimes adopted. In our case, the pth sample quantile $Q_X(p)$ is equal to $X_{(np)}$ if np is an integer, and $X_{([np+1])}$ if np is not an integer.

A point estimate is not sufficient for inference purposes. We know from Theorem 10.1 of Chapter 2 that the rth-order statistic is a consistent estimator of the pth quantile of a distribution when $n \to \infty$ and $r/n \to p$. However, consistency is only a large-sample property. We would like a procedure for interval estimation of κ_p which will enable us to attach a confidence coefficient to our estimate for the given (finite) sample size. A logical choice for the confidence interval endpoints are two order statistics, say $X_{(r)}$ and $X_{(s)}, r < s$, from the random sample drawn from the population F_X. To find the $100(1-\alpha)\%$ confidence interval, we must then find the two integers r and $s, 1 \leqslant r < s \leqslant n$, such that

$$P(X_{(r)} < \kappa_p < X_{(s)}) = 1 - \alpha$$

for some given number $0 < \alpha < 1$. The quantity $1 - \alpha$, which we frequently denote by γ, is called the *confidence level* or the *confidence coefficient*. Now the event $X_{(r)} < \kappa_p$ occurs if and only if either $X_{(r)} < \kappa_p < X_{(s)}$ or $\kappa_p > X_{(s)}$, and these latter two events are clearly mutually exclusive. Therefore, for all $r < s$,

$$P(X_{(r)} < \kappa_p) = P(X_{(r)} < \kappa_p < X_{(s)}) + P(\kappa_p > X_{(s)})$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

or, equivalently,

$$P(X_{(r)} < \kappa_p < X_{(s)}) = P(X_{(r)} < \kappa_p) - P(X_{(s)} < \kappa_p) \tag{2.1}$$

Since we assumed that F_X is a strictly increasing function,

$$X_{(r)} < \kappa_p \qquad \text{if and only if } F_X(X_{(r)}) < F_X(\kappa_p) = p$$

But when F_X is continuous, the PIT implies that the probability distribution of the random variable $F_X(X_{(r)})$ is the same as that of $U_{(r)}$, the rth-order statistic from the uniform distribution over the interval (0,1). Further, since $F_X(\kappa_p) = p$ by the definition of κ_p, we have

$$\begin{aligned} P(X_{(r)} < \kappa_p) &= P[F_X(X_{(r)}) < p] \\ &= \int_0^p \frac{n!}{(r-1)!(n-r)!} x^{r-1}(1-x)^{n-r}\,dx \end{aligned} \tag{2.2}$$

Thus, while the distribution of the rth-order statistic depends on the population distribution F_X, the probability in (2.2) does not. A confidence-interval procedure based on (2.1) is therefore distribution free.

In order to find the interval estimate of κ_p, substitution of (2.2) back into (2.1) indicates that r and s should be chosen such that

$$\begin{aligned} &\int_0^p n\binom{n-1}{r-1} x^{r-1}(1-x)^{n-r}\,dx \\ &\quad - \int_0^p n\binom{n-1}{s-1} x^{s-1}(1-x)^{n-s}\,dx = 1-\alpha \end{aligned} \tag{2.3}$$

Clearly, this one equation will not give a unique solution for the two unknowns, r and s, and additional conditions are needed. For example, if we want the narrowest possible interval for a fixed confidence coefficient, r and s should be chosen such that (2.3) is satisfied and $X_{(s)} - X_{(r)}$, or $E[X_{(s)} - X_{(r)}]$, is as small as possible. Alternatively, we could minimize $s - r$.

The integrals in (2.2) or (2.3) can be evaluated by integration by parts or by using tables of the incomplete beta function. However, (2.2) can be expressed in another form after integration by parts as follows:

$$\begin{aligned} P(X_{(r)} < \kappa_p) &= \int_0^p n\binom{n-1}{r-1} x^{r-1}(1-x)^{n-r}\,dx \\ &= n\binom{n-1}{r-1}\left[\frac{x^r}{r}(1-x)^{n-r}\Big|_0^p + \frac{n-r}{r}\int_0^p x^r(1-x)^{n-r-1}\,dx\right] \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$
\begin{aligned}
&= \binom{n}{r} p^r (1-p)^{n-r} + n\binom{n-1}{r}\left[\frac{x^{r+1}}{r+1}(1-x)^{n-r-1}\Big|_0^p \right.\\
&\quad \left. + \left(\frac{n-r-1}{r+1}\right)\int_0^p x^{r+1}(1-x)^{n-r-2}\,dx\right]\\
&= \binom{n}{r} p^r (1-p)^{n-r} + \binom{n}{r+1} p^{r+1}(1-p)^{n-r-1}\\
&\quad + n\binom{n-1}{r+1}\int_0^p x^{r+1}(1-x)^{n-r-2}\,dx
\end{aligned}
$$

After repeating this integration by parts $n - r$ times, the result will be

$$
\begin{aligned}
&\binom{n}{r} p^r (1-p)^{n-r} + \binom{n}{r+1} p^{r+1}(1-p)^{n-r-1} + \cdots\\
&\quad + \binom{n}{n-1} p^{n-1}(1-p) + n\binom{n-1}{n-1}\int_0^p x^{n-1}(1-x)^0\,dx\\
&\quad = \sum_{j=0}^{n-r}\binom{n}{r+j} p^{r+j}(1-p)^{n-r-j}
\end{aligned}
$$

or, after substituting $r + j = i$,

$$
P(X_{(r)} < \kappa_p) = \sum_{i=r}^{n}\binom{n}{i} p^i (1-p)^{n-i} \tag{2.4}
$$

In this final form, the integral in (2.2) is expressed as the sum of the last $n - r + 1$ terms of the binomial distribution with parameters n and p. Thus, the probability in (2.1) can be expressed as

$$
\begin{aligned}
P(X_{(r)} < \kappa_p < X_{(s)}) &= \sum_{i=r}^{n}\binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=s}^{n}\binom{n}{i} p^i (1-p)^{n-i}\\
&= \sum_{i=r}^{s-1}\binom{n}{i} p^i (1-p)^{n-i}\\
&= P(r \leqslant K \leqslant s-1)
\end{aligned} \tag{2.5}
$$

where K has a binomial distribution with parameters n and p. This form is probably the easiest to use in choosing r and s such that $s - r$ is a minimum for fixed α. Note that from (2.5) it is clear that this probability does not depend on the underlying cdf as long as it is

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

continuous. The resulting confidence interval is therefore distribution free.

In order to find the confidence interval for κ_p based on two-order statistics, the right-hand side of (2.5) is set equal to $1-\alpha$ and the search for r and s is begun. Because of the discreteness of the binomial distribution, the exact nominal confidence level frequently cannot be achieved. In such cases, the confidence level requirement can be changed from "equal to" to "at least equal to" $1-\alpha$. We usually let $\gamma \geqslant 1-\alpha$ denote the *exact confidence level*.

The result obtained in (2.4) found by integration of (2.2) can also be obtained by arguing as follows. This argument, based on simple counting, is used frequently in the context of various nonparametric procedures where order statistics are involved. Note that for any p, the event $X_{(r)} < \kappa_p$ occurs if and only if at least r of the n sample values, $X_1, X_2, \ldots, X_n$, are less than κ_p. Thus

$$\begin{aligned} P(X_{(r)} < \kappa_p) = {} & P(\text{exactly } r \text{ of the } n \text{ observations are } < \kappa_p) \\ & + P(\text{exactly } (r+1) \text{ of the n observations} \\ & \quad \text{are } < \kappa_p) + \cdots \\ & + P(\text{exactly } n \text{ of the } n \text{ observations are } < \kappa_p) \end{aligned}$$

In other words,

$$P(X_{(r)} < \kappa_p) = \sum_{i=r}^{n} P(\text{exactly } i \text{ of the } n \text{ observations are } < \kappa_p)$$

This is a key observation. Now, the probability that exactly i of the n observations are less than κ_p can be found as the probability of i successes in n independent Bernoulli trials, since the sample observations are all independent and each observation can be classified either as a success or a failure, where a success is defined as any observation being less than κ_p. The probability of a success is $P(X_i < \kappa_p) = p$. Thus, the required probability is given by the binomial probability with parameters n and p. In other words,

$$P(\text{exactly } i \text{ of the } n \text{ sample values are } < \kappa_p) = \binom{n}{i} p^i (1-p)^{n-i}$$

and therefore

$$P(X_{(r)} < \kappa_p) = \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

This completes the proof.

In summary, the $(1-\alpha)100\%$ confidence interval for the pth quantile is given by $(X_{(r)}, X_{(s)})$, where r and s are integers such that $1 \leqslant r < s \leqslant n$ and

$$P(X_{(r)} < \kappa_p < X_{(s)}) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \geqslant 1 - \alpha \tag{2.6}$$

As indicated earlier, without a second condition, the confidence interval endpoints will not be unique. One common approach in this case is to assign the probability $\alpha/2$ in each (right and left) tail. This yields the so-called "equal-tails" interval, where r and s are the *largest* and *smallest* integers $(1 \leqslant r < s \leqslant n)$ respectively such that

$$\sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \leqslant \frac{\alpha}{2} \qquad \text{and} \qquad \sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \geqslant 1 - \frac{\alpha}{2} \tag{2.7}$$

respectively. These equations are easy to use in conjunction with Table C of the Appendix, where cumulative binomial probabilities are given. The exact confidence level is found from Table C as

$$\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} = \gamma \tag{2.8}$$

If the sample size is larger than 20 and therefore beyond the range of Table C, we can use the normal approximation to the binomial distribution with a continuity correction. The solutions are

$$r = np + 0.5 - z_{\alpha/2}\sqrt{np(1-p)} \quad \text{and} \quad s = np + 0.5 + z_{\alpha/2}\sqrt{np(1-p)} \tag{2.9}$$

where $z_{\alpha/2}$ satisfies $\Phi(z_{\alpha/2}) = 1 - \alpha/2$, as defined in Chapter 3. We round the result in (2.9) down to the nearest integer for r and round up for s in order to be conservative (or to make the confidence level at least $1-\alpha$).

Example 2.1 Suppose $n = 10, p = 0.35$, and $1 - \alpha = 0.95$. Using (2.7) with Table C shows that $r - 1 = 0$ and $s - 1 = 7$, making $r = 1$ and

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

$s = 8$. The confidence interval for the 0.35th quantile is $(X_{(1)}, X_{(8)})$ with exact confidence level from (2.8) equal to $0.9952 - 0.0135 = 0.9817$. The normal approximation gives $r = 1$ and $s = 7$ with approximate confidence level 0.95.

Now suppose that $n = 10, p = 0.10$, and $1 - \alpha = 0.95$. Table C shows that $s - 1 = 3$ and no value of $r - 1$ satisfies the left-hand condition of (2.7) so we take the smallest possible value $r - 1 = 0$. The confidence interval for the 0.10th quantile is then $(X_{(1)}, X_{(4)})$ with exact confidence $0.9872 - 0 = 0.9872$.

Another possibility is to find those values of r and s such that $s - r$ is a minimum. This requires a trial-and-error solution in making (2.8) at least $1 - \alpha$. In the two situations described in Example 2.1, this approach yields the same values of r and s as the equal-tails approach. But if $n = 11, p = 0.25$ and $1 - \alpha = 0.95$, (2.7) gives $r = 0$ and $s = 7$ with exact confidence coefficient 0.9924 from (2.8). The values of r and s that make $s - r$ as small as possible and make (2.8) at least 0.95 are $r = 0$ and $s = 6$, with exact confidence coefficient 0.9657. The reader can verify these results.

5.3 HYPOTHESIS TESTING FOR A POPULATION QUANTILE

In a hypothesis testing type of inference concerned with quantiles, a distribution-free procedure is also possible. Given the order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ from any *unspecified* but continuous distribution F_X, a null hypothesis concerning the value of the pth quantile is written

$$H_0\colon \kappa_p = \kappa_p^0$$

where κ_p^0 and p are both specified numbers. Under H_0, since κ_p^0 is the pth quantile of F_X, we have, by definition, $P(X \leqslant \kappa_p^0) = p$ and therefore we expect about np of the sample observations to be smaller than κ_p^0 if H_0 is true. If the actual number of sample observations smaller than κ_p^0 is considerably smaller than np, the data suggest that the true pth quantile is larger than κ_p^0 or there is evidence against H_0 in favor of the one-sided upper-tailed alternative

$$H_1\colon \kappa_p > \kappa_p^0$$

This implies it is reasonable to reject H_0 in favor of H_1 if at most $r - 1$ sample observations are smaller than κ_p^0, for some r. Now if at most $r - 1$ sample observations are smaller than κ_p^0, then it must be true

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

that the rth-order statistic $X_{(r)}$ in the sample satisfies $X_{(r)} > \kappa_p^0$. Therefore an appropriate rejection region R is

$$X_{(r)} \in R \qquad \text{for } X_{(r)} > \kappa_p^0 \tag{3.1}$$

For a specified significance level α, the integer r should be chosen such that

$$P(X_{(r)} > \kappa_p^0 \mid H_0) = 1 - P(X_{(r)} \leqslant \kappa_p^0 \mid H_0) \leqslant \alpha$$

or, using (2.4), r is the largest integer such that

$$1 - \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \leqslant \alpha \tag{3.2}$$

We now express the rejection region in another form in order to be consistent with our later presentation in Section 5.4 for the sign test. Note that $X_{(r)} > \kappa_p^0$ if and only if at most $r-1$ of the observations are less than κ_p^0, so that at least $n-(r-1) = n-r+1$ of the observations are greater than κ_p^0. Define the random variable K as the total number of plus signs among the n differences $X_{(i)} - \kappa_p^0$ (the number of positive differences). Then the rejection region in (3.1) can be equivalent stated as

$$K \in R \qquad \text{for } K \geqslant n-r+1$$

The differences $X_i - \kappa_p^0$, $i = 1, 2, \ldots, n$, are independent random variables, each having either a plus or a minus sign, and the probability of a plus sign under H_0 is

$$P(X_i - \kappa_p^0 > 0) = P(X_i > \kappa_p^0) = 1 - p$$

Hence, since K is the number of plus signs, we can write $K = \sum_{i=1}^{n} I(X_i > \kappa_p^0)$ where $I(A) = 1$ when the event A occurs and is 0 otherwise. From the preceding discussion, the indicator variables $I(X_i > \kappa_p^0)$, $i = 1, 2, \ldots, n$, are independent Bernoulli random variables with probability of success $1-p$ under H_0. Thus under H_0, the distribution of K is binomial with parameters n and $1-p$ and so r must be chosen to satisfy

$$P(K \geqslant n-r+1 \mid H_0) = \sum_{i=n-r+1}^{n} \binom{n}{i} (1-p)^i p^{n-i} \leqslant \alpha \tag{3.3}$$

which can be shown to agree with the statement in (3.2), by a change of summation index from i to $n-i$. The advantage of using (3.2) is that

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

cumulative binomial probabilities are directly involved and these are given in Table C.

On the other hand, if many more than np observations are smaller than κ_p^0, there is support against H_0 in favor of the one-sided lower-tailed alternative H_1: $\kappa_p < \kappa_p^0$. Then we should reject H_0 if the number of sample observations smaller than κ_p^0 is at least, say s. This leads to the rejection region

$$X_{(s)} \in R \qquad \text{for } X_{(s)} < \kappa_p^0$$

but this is equivalent to saying that the number of observations larger than κ_p^0 must be at most $n - s$. Thus, based on the statistic K, defined before as the number of positive differences, the appropriate rejection region for the one-sided lower-tailed alternative H_1: $\kappa_p < \kappa_p^0$ is

$$K \in R \qquad \text{for } K \leqslant n - s$$

where s is the largest integer such that

$$P(K \leqslant n - s | H_0) = \sum_{i=0}^{n-s} \binom{n}{i} (1-p)^i p^{n-i} \leqslant \alpha \tag{3.4}$$

For the two-sided alternative H_1: $\kappa_p \neq \kappa_p^0$, the rejection region consists of the union of the two pieces specified above,

$$K \in R \qquad \text{for } K \leqslant n - s \text{ or } K \geqslant n - r + 1 \tag{3.5}$$

where r and s are integers such that each of (3.2) and (3.4) is less than or equal to $\alpha/2$.

Note that Table C can be used to find the exact critical values for $n \leqslant 20$, where $\theta = p$ in (3.2) and $\theta = 1 - p$ in (3.4). For example sizes larger than 20 the normal approximation to the binomial distribution with a continuity correction can be used. The rejection region for H_1: $\kappa_p > \kappa_p^0$ is

$$K \geqslant 0.5 + n(1-p) + z_\alpha \sqrt{np(1-p)}$$

For H_1: $\kappa_p < \kappa_p^0$, the rejection region is

$$K \leqslant -0.5 + n(1-p) - z_\alpha \sqrt{np(1-p)}$$

The rejection region for H_1: $\kappa_p \neq \kappa_p^0$ is the combination of these two with z_α replaced by $z_{\alpha/2}$. Note that in all these formulas the standard normal deviate, say z_b, is such that the area to the right is b; in other words, z_b is the $100(1-b)$th percentile [or the $(1-b)$th quantile] of the standard normal distribution.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 3.1 (p. 167) summarizes the appropriate rejection regions for the quantile test and the corresponding P values, both exact and approximate, where K_0 is the observed value of the statistic K, the number of positive differences.

Example 3.1 The Educational Testing Service reports that the 75th percentile for scores on the quantitative portion of the Graduate Record Examination (GRE) is 693 in a certain year. A random sample of 15 first-year graduate students majoring in statistics report their GRE quantitative scores as 690, 750, 680, 700, 660, 710, 720, 730, 650, 670, 740, 730, 660, 750, and 690. Are the scores of students majoring in statistics consistent with the 75th percentile value for this year?

Solution The question in this example can be answered either by a hypothesis testing or a confidence interval approach. We illustrate both approaches at the 0.05 level. Here we are interested in the 0.75th quantile (the third quartile) so that $p = 0.75$, and the hypothesized value of the 0.75th quantile, $\kappa_{0.75}^{0}$, is 693. Thus, the null hypothesis H_0: $\kappa_{0.75} = 693$ is to be tested against a two-sided alternative H_1: $\kappa_{0.75} \neq 693$. The value of the test statistic is $K = 8$, since there are eight positive differences among $X_i - 693$, and the two-sided rejection region is $K \in R$ for $K \leqslant n - s$ or $K \geqslant n - r + 1$, where r and s are the largest integers that satisfy (3.2) and (3.4) with $\alpha/2 = 0.025$. For $n = 15, p = 0.75$, Table C shows that 0.0173 is the largest left-tail probability that does not exceed 0.025, so $r - 1 = 7$ and hence $r = 8$; similarly, 0.0134 is the largest left-tail probability that does not exceed 0.025 for $n = 15$ and $1 - p = 0.25$ (note the change in the success probability) so that $n - s = 0$ and $s = 15$. The two-sided critical region then is $K \leqslant 0$ or $K \geqslant 8$, and the exact significance level for this distribution-free test is $(0.0134 + 0.0173) = 0.0307$. Since the observed $K = 8$ falls in this rejection region, there is evidence that for this year, the scores for the graduate majors in statistics are not consistent with the reported 75th percentile for all students in this year.

In order to find the P value, note that the alternative is two-sided and so we need to find the two one-tailed probabilities first. Using Table C with $n = 15$ and $\theta = 0.25$ we find $P(K \leqslant 8|H_0) = 0.9958$ and $P(K \geqslant 8|H_0) = 1 - 0.9827 = 0.0173$. Taking the smaller of these two values and multiplying by 2, the required P value is 0.0346, which also suggests rejecting the null hypothesis.

In order to find a 95% confidence interval for $\kappa_{0.75}$, we use (2.7). For the lower index r, the inequality on the left applies. From Table C with $n = 15$ and $\theta = 0.75$, the largest value of x such that

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 3.1 Hypothesis testing guide for quantiles

Alternative	*Rejection region*	*P value*
$\kappa_p > \kappa_p^0$	*Exact* $X_{(r)} > \kappa_p^0$ or $K \geqslant n - r + 1$, r from (3.2)	*Exact* $P_U = \sum_{k=K_O}^{n} \binom{n}{k} (1-p)^k p^{n-k}$
	Approximate $K \geqslant 0.5 + n(1-p)$ $+ z_\alpha \sqrt{n(1-p)p}$	*Approximate* $P_U^* = 1 - \Phi\left(\frac{K_0 - 0.5 - n(1-p)}{\sqrt{np(1-p)}}\right)$
$\kappa_p > \kappa_p^0$	*Exact* $X_{(s)} < \kappa_p^0$ or $K \leqslant n - s$, s from (3.4)	*Exact* $P_L = \sum_{k=0}^{K_O} \binom{n}{k} (1-p)^k p^{n-k}$
	Approximate $K \leqslant -0.5 + n(1-p)$ $- z_\alpha \sqrt{n(1-p)p}$	*Approximate* $P_L^* = \Phi\left(\frac{K_0 + 0.5 - n(1-p)}{\sqrt{np(1-p)}}\right)$
$\kappa_p \neq \kappa_p^0$	*Exact* $X_{(r)} > \kappa_p^0$ or $X_{(s)} < \kappa_p^0$ or $K \geqslant n - r + 1$ or $K \leqslant n - s$, r and s from (3.5)	*Exact* 2 min (P_U, P_L)
	Approximate $K \geqslant 0.5 + n(1-p)$ $+ z_{\alpha/2} \sqrt{n(1-p)p}$ or $K \leqslant -0.5 + n(1-p)$ $- z_{\alpha/2} \sqrt{n(1-p)p}$	*Approximate* 2 min (P_U^*, P_L^*)

the cumulative probability is less than or equal to 0.025 is 7, which yields $r = 8$ with corresponding probability 0.0173. For the upper index s, the inequality on the right in (2.7) applies, again with $n = 15$ and $\theta = 0.75$. From Table C, the smallest value of x such that the cumulative probability is greater than or equal to 0.975 is 14, so that $s = 15$ with corresponding probability 0.9866. The desired 95% confidence interval endpoints are $X_{(8)}$ and $X_{(15)}$ which are 700 and 750, respectively. The exact confidence level using (2.8) is $\gamma = 0.9866 - 0.0173 = 0.9693$. Thus we have at least 95% confidence, or exactly 96.93% confidence, that the 75th percentile (or the 0.75th quantile) score of students majoring in statistics lies somewhere between 700 and 750. Note that, on the basis of this confidence interval we would again reject H_0: $\kappa_{0.75} = 693$ in favor of the alternative H_1: $\kappa_{0.75} \neq 693$, since the hypothesized value of the 75th percentile lies outside of the confidence interval.

One of the special quantiles of a distribution is the median (the 0.5th quantile or the 50th percentile). The median is an important and useful parameter in many situations, particularly when the underlying distribution is skewed. This is mainly because the median is a far more robust estimate of the center of a distribution than the mean. Quantile tests and confidence intervals discussed earlier can both be applied to the case of the median with $p = 0.5$. However, because of its special importance, the case for the median is treated separately in the next section.

5.4 THE SIGN TEST AND CONFIDENCE INTERVAL FOR THE MEDIAN

Suppose that a random sample of N observations $X_1, X_2, \ldots, X_N$ is drawn from a population F_X with an unknown median M, where F_X is assumed to be continuous and strictly increasing, at least in the vicinity of M. In other words, the N observations are independent and identically distributed, and $F_X^{-1}(0.5) = M$, uniquely. The total sample size notation is changed from n to N in this section in order to be consistent with the notation in the rest of this book.

The hypothesis to be tested concerns the value of the population median

$$H_0: M = M_0$$

where M_0 is a specified value, against a corresponding one- or two-sided alternative. Since by assumption F_X has a unique median, the null hypothesis states that M_0 is that value of X which divides the area

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

under the pdf into two equal parts. An equivalent symbolic representation of H_0 is

$$H_0\colon \theta = P(X > M_0) = P(X < M_0) = 0.50$$

Recalling the arguments used in developing a distribution-free test for an arbitrary quantile, we note that if the sample data are consistent with the hypothesized median value, on the average half of the sample observations will lie above M_0 and half below. Thus the number of observations larger than M_0, denoted by K, can be used to test the validity of the null hypothesis. Also, when the sample observations are dichotomized in this way, they constitute a set of n independent random variables from the Bernoulli population with parameter $\theta = P(X > M_0)$, regardless of the population F_X. The sampling distribution of the random variable K then is the binomial probability distribution with parameters N and θ, and θ equals 0.5 if the null hypothesis is true. Since K is actually the number of plus signs among the N differences $X_i - M_0$, $i = 1, 2, \ldots, N$, the nonparametric test based on K is called the *sign test*.

The rejection region for the upper-tailed alternative

$$H_1\colon M > M_0 \qquad \text{or} \qquad \theta = P(X > M_0) > P(X < M_0)$$

is

$$K \in R \quad \text{for } K \geqslant k_\alpha$$

where k_α is chosen to be the smallest integer which satisfies

$$P(K \geqslant k_\alpha \mid H_0) = \sum_{i=k_\alpha}^{N} \binom{N}{i} (0.5)^N \leqslant \alpha \tag{4.1}$$

Any table of the binomial distribution, like Table C of the Appendix, can be used with $\theta = 0.5$ to find the particular value of k_α for the given N and α, but Table G of the Appendix is easier to use because it gives probabilities in both tails. Similarly, for a one-sided test with the lower-tailed alternative

$$H_1\colon M < M_0 \qquad \text{or} \qquad \theta = P(X > M_0) < P(X < M_0)$$

the rejection region for an α-level test is

$$K \in R \qquad \text{for } K \leqslant k'_\alpha$$

where k'_α is the largest integer satisfying

$$\sum_{i=0}^{k'_\alpha} \binom{N}{i} (0.5)^N \leqslant \alpha \tag{4.2}$$

If the alternative is two-sided,

$$H_1\colon M \neq M_0 \qquad \text{or} \qquad \theta = P(X > M_0) \neq P(X < M_0)$$

the rejection region is $K \geqslant k_{\alpha/2}$ or $K \leqslant k'_{\alpha/2}$, where $k_{\alpha/2}$ and $k'_{\alpha/2}$ are respectively, the smallest and the largest integers such that

$$\sum_{i=k_{\alpha/2}}^{N} \binom{N}{i} (0.5)^N \leqslant \frac{\alpha}{2} \qquad \text{and} \qquad \sum_{i=0}^{k'_{\alpha/2}} \binom{N}{i} (0.5)^N \leqslant \frac{\alpha}{2} \tag{4.3}$$

Obviously, we have the relation $k_{\alpha/2} = N - k'_{\alpha/2}$.

The sign test statistics with these rejection regions are consistent against the respective one- and two-sided alternatives. This is easy to show by applying the criterion of consistency given in Chapter 1. Since $E(K/N) = \theta$ and $\text{var}(K/N) = \theta(1-\theta)/N \to \infty$ as $N \to \infty, K$ provides a consistent test statistic.

P VALUE

The P value expressions for the sign test can be obtained as in the case of a general quantile test with $p = 0.5$. The reader is referred to Table 3.1, with n replaced by N throughout. For example, if the alternative is upper-tailed, $H_1\colon M > M_0$, and K_O is the observed value of the sign statistic, the P value for the sign test is given by the binomial probability in the upper-tail

$$\sum_{i=K_O}^{N} \binom{N}{i} (0.5)^N$$

This value is easily read as a right-tail probability from Table G for the given N.

NORMAL APPROXIMATIONS

We could easily generate tables to apply the exact sign test for any sample size N. However, we know that the normal approximation to the binomial is especially good when $\theta = 0.50$. Therefore, for moderate values of N (say at least 12), the normal approximation to the binomial can be used to determine the rejection regions. Since this is a continuous approximation to a discrete distribution, a continuity correction of 0.5 may be incorporated in the calculations. For example, for the alternative H_1: $M > M_0, H_0$ is rejected for $K \geqslant k_\alpha$, where k_α satisfies

$$k_\alpha = 0.5N + 0.5 + 0.5\sqrt{N}z_\alpha \tag{4.4}$$

Similarly, the approximate P value is

$$1 - \Phi\left(\frac{K_O - 0.5 - 0.5N}{\sqrt{0.25N}}\right) \tag{4.5}$$

ZERO DIFFERENCES

A zero difference arises whenever $X_i = M_0$ for at least one i. Theoretically, zero differences do not cause a problem because the population was assumed to be continuous in the vicinity of the median. In reality, of course, zero differences can and do occur, either because the assumption of continuity is in error or because of imprecise measurements. Many zeros can be avoided by taking measurements to a larger number of significant figures.

The most common treatment of zeros is simply to ignore them and reduce N accordingly. The inferences are then conditional on the observed number of nonzero differences. An alternative approach is to treat half of the zeros as plus and half as minus. Another possibility is to assign to all that sign which is least conducive to rejection of H_0; this is a strictly conservative approach. Finally, we could let chance determine the signs of the zeros by, say, flipping a balanced coin. These procedures are compared in Putter (1955) and Emerson and Simon (1979). A complete discussion, including more details on P values, is given in Pratt and Gibbons (1981). Randles (2001) proposed a more conservative method of handling zeros.

POWER FUNCTION

In order to calculate the power of any test, the distribution of the test statistic under the alternative hypothesis should be available in a reasonably tractable form. In contrast to most nonparametric tests, the power function of the quantile tests is simple to determine since, in general, the random variable K follows the binomial probability distribution with parameters N and θ, where, for the pth quantile, $\theta = P(X_i > \kappa_p)$. For the sign test the quantile of interest is the median and $\theta = P(X_i > M_0)$. For illustration, we will only consider the power of the sign test against the one-sided upper-tailed alternative H_1: $M > M_0$.

The power of the test is a function of the unknown parameter θ, and the power curve or the power function is a graph of power versus various values of θ, under the alternative. By definition, the power of the sign test against the alternative H_1 is the probability

$$\mathrm{Pw}(\theta) = P(K \geqslant k_\alpha | H_1)$$

Under H_1, the distribution of K is binomial with parameters N and $\theta = P(X_i > M_0 \mid H_1)$ so the expression for power can be written as

$$\mathrm{Pw}(\theta) = \sum_{i=k_\alpha}^{N} \binom{N}{i} \theta^i (1-\theta)^{N-i}$$

where k_α is the smallest integer such that

$$\sum_{i=k_\alpha}^{N} \binom{N}{i} (0.5)^N \leqslant \alpha$$

Thus, in order to evaluate the power function for the sign test, we first need to find the critical value k_α for a given level α, say 0.05. Then we need to calculate the probability $\theta = P(X_i > M_0 | H_1)$. If the power function is desired for a more parametric type of situation where the population distribution is fully specified then θ can be calculated. Such a power function would be desirable for comparisons between the sign test and some parametric test for location.

As an example, we calculate the power of the sign test of H_0: $M = 28$ versus H_1: $M > 28$ for $N = 16$ at a significance level 0.05, under the assumption that the population is normally distributed with standard deviation 1 and the median is $M = 29.04$. Table G shows that the rejection region at $\alpha = 0.05$ is $K \geqslant 12$ so that $k_\alpha = 12$ and the exact size of this sign test is 0.0384. Now, under the assumptions given, we can evaluate the underlying probability of a success θ as

$$\begin{aligned}\theta &= P(X > 28 | H_1) \\ &= P\left(\frac{X - 29.04}{1} > \frac{28 - 29.04}{1}\right) \\ &= P(Z > -1.04) \\ &= 1 - \Phi(-1.04) = 0.8505 \\ &= 0.85, \text{ say}\end{aligned}$$

Note that the value of θ is larger than 0.5, which is in the legitimate region of the alternative H_1. Thus,

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$\text{Pw}(0.85) = \sum_{i=12}^{16} \binom{16}{i} (0.85)^i (0.15)^{16-i}$$
$$= 1 - \sum_{i=0}^{11} \binom{16}{i} (0.85)^i (0.15)^{16-i} = 0.9209$$

This would be directly comparable with the normal theory test of H_0: $\mu = 28$ versus H_1: $\mu = 29.04$, say with $\sigma = 1$, since the mean and median coincide for the normal distributions. The rejection region for this parametric test with $\alpha = 0.05$ is $\bar{X} > 28 + z_{0.05}/\sqrt{16} = 28.41$, and the power is

$$\text{Pw}(29.04) = P[\bar{X} > 28.41 | X \sim \text{normal}(29.04, 1)]$$
$$= P\left(\frac{\bar{X} - 29.04}{1/\sqrt{16}} > \frac{28 - 29.04}{0.25}\right)$$
$$= P(Z > -2.52)$$
$$= 0.9941$$

Thus, the power of the normal theory test is larger than the power of the sign test, which is of course expected, since the normal theory test is known to be the best test when the population is normal. The problem with a direct comparison of the exact sign test with the normal theory test is that the powers of any two tests are comparable only when their sizes or significance levels are the same or nearly the same. In our case, the sign test has an exact size of 0.0384 whereas the normal theory test has exact size 0.05. This increase in the size of the test inherently biases the power comparison in favor of the normal theory test.

In order to ensure a more fair comparison, we might make the exact size of the sign test equal to 0.05 by using a randomized version of the sign test (as explained in Chapter 1). Alternatively, we might find the normal theory test of size $\alpha = 0.0384$ and compare the power of that test with the sign-test power of 0.9209. In this case, the rejection region is $\bar{X} > 28 + z_{0.0384}/\sqrt{16} = 28.44$ and the power is $\text{Pw}(29.04) = 0.9918$. This is still larger than the power of the sign test at $\alpha = 0.0384$ but two comments are in order. First and foremost, we have to assume that the underlying distribution is normal to justify using the normal theory test. No such assumption is necessary for the sign test. If the sample size N is larger, the calculated power is an approximation to the power of the normal theory test, by the central limit theorem. However, for the sign test, the size and the power calculations can be made exactly for all sample sizes and no distribution assumptions are needed other than continuity. Further, the normal theory test is affected by the assumption about the population standard deviation σ, whereas the sign test calculations do not demand such knowledge. In order to obtain the power function, we can calculate the power at several values of M in the alternative region ($M > 28$) and then plot the power versus the values of the median. This is easier under the normal approximation and is shown below.

Since under the alternative hypothesis H_1, the sign test statistic K has a binomial distribution with parameters N and $\theta = P(X > M_0|H_1)$, and the binomial distribution can be well approximated by the normal distribution, we can derive expressions to approximate the power of the sign test based on the normal approximation. These formulas are useful in practice for larger sample sizes and/or θ values for which exact tables are unavailable, although this appears to be much less of a problem with currently available software. We consider the one-sided upper-tailed case H_1: $M_1 > M_0$ for illustration; approximate power expressions for the other cases are left as exercises for the reader. The power for this alternative can be evaluated using the normal approximation with a continuity correction as

$$\begin{aligned}
\mathrm{Pw}(M_1) &= P(K \geqslant k_\alpha | H_1\colon M_1 > M_0) \\
&= P\left(Z > \frac{k_\alpha - N\theta - 0.5}{\sqrt{N\theta(1-\theta)}}\right) \\
&= 1 - \Phi\left(\frac{k_\alpha - N\theta - 0.5}{\sqrt{N\theta(1-\theta)}}\right) \qquad (4.6)
\end{aligned}$$

where $\theta = P(X > M_1 | M_1 > M_0)$ and k_α is such that

$$\begin{aligned}
\alpha &= P(K \geqslant k_\alpha | H_0) \\
&= P\left(Z > \frac{k_\alpha - N/2 - 0.5}{\sqrt{N/4}}\right) \\
&= 1 - \Phi\left(\frac{2k_\alpha - N - 1}{\sqrt{N}}\right) \qquad (4.7)
\end{aligned}$$

The equality in (4.7) implies that $k_\alpha = [N + 1 + \sqrt{N}\Phi^{-1}(1-\alpha)]/2$. Substituting this back into (4.6) and simplifying gives

Table 4.1 Normal approximation to power of the sign test for the median when $N = 16$

θ	0.5	0.55	0.6	0.65	0.70	0.75	0.80	0.85	0.90
Power	0.0461	0.0918	0.1629	0.2639	0.3960	0.5546	0.7255	0.8802	0.9773

$$\begin{aligned}\text{Pw}(M_1) &= P\left\{Z > \frac{0.5[N+1+\sqrt{N}\Phi^{-1}(1-\alpha)] - N\theta - 0.5}{\sqrt{N\theta(1-\theta)}}\right\} \\ &= 1 - \Phi\left[\frac{N(0.5-\theta)+0.5\sqrt{N}z_\alpha}{\sqrt{N\theta(1-\theta)}}\right] \end{aligned} \tag{4.8}$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$th quantile of the standard normal distribution. For example, $z_{0.05} = 1.645$ and $z_{0.85} = -1.04$. Note that $z_\alpha = -z_{1-\alpha}$. The approximate power values are calculated and shown in Table 4.1 for $N = 16$ and $\alpha = 0.05$. A graph of the power function is shown in Figure 4.1.

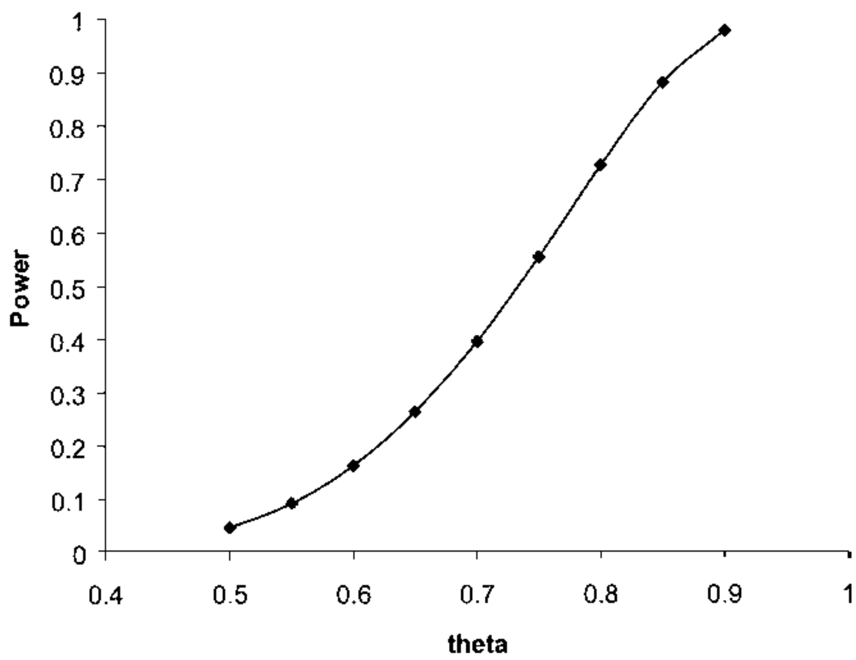


Fig. 4.1 Normal approximation to the power function of the sign test for the median.

It should be noted that the power of the sign test depends on the alternative hypothesis through the probability $\theta = P(X > M_0|H_1: M_1 > M_0)$. Under H_0, we have $\theta = 0.5$, whereas $\theta > 0.5$ under H_1, since if $M_1 > M_0$,

$$P(X > M_0|H_1: M = M_1 > M_0) > P(X > M_1|H_1: M = M_1 > M_0)$$

and therefore $\theta = P(X > M_0|H_1) > P(X > M_1|H_1) = 0.5$. Thus, the power of the sign test depends on the "distance" between the values of θ under the null hypothesis (0.5) and under the alternative and specification of a value of $\theta > 0.5$ is necessary for the power calculation. Noether (1987) suggested choosing a value of θ based on past information or a pilot study, or based on an "odds-ratio." In the normal theory test (such as the *t* test), however, the power depends directly on the "distance" $M_1 - M_0$, the values of the median under the null hypothesis and under the alternative. Note also that the approximate power is exactly equal to the nominal size of the test when $\theta = 0.5$ (i.e., the null hypothesis is true). Expressions for approximate power against other alternatives are left as exercises for the reader.

SIMULATED POWER

The power function for the sign test is easily found, particularly when the normal approximation is used for calculations. For many other nonparametric tests, however, the power function can be quite difficult to calculate. In such cases, computer simulations can be used to estimate the power. Here we use a MINITAB Macro program to simulate the power of the sign test when the underlying distribution is normal with mean = median = M and variance σ^2. The null hypothesis is H_0: $M = M_0$ and the alternative is H_0: $M = M_1 > M_0$. First we need to find the relationship between M_0, M_1 and θ. Recall that $\theta = P(X_i > M_0|H_1)$, so assuming X is normally distributed with variance σ^2, we get

$$\begin{aligned}\theta &= P\left(\frac{X - M_1}{\sigma} > \frac{M_0 - M_1}{\sigma}\right)\\ &= \Phi\left(\frac{M_1 - M_0}{\sigma}\right)\end{aligned}$$

This gives $(M_1 - M_0)/\sigma = \Phi^{-1}(\theta)$. Now let us assume arbitrarily that $M_0 = 0.5$ and $\sigma^2 = 1$. Then if $\theta = 0.55$, say, $\Phi^{-1}(0.55) = 0.1256$ and

$M_1 = 0.6256$. Next we need to specify a sample size and probability of a type I error for the test. We arbitrarily choose $N = 13$ and $\alpha = 0.05$. From Table G, 0.0461 is closest to 0.05 and this gives a test with rejection region $K \geqslant 10$ for exact size 0.0461.

First we generate 1000 random samples, each of size 13, from a normal distribution with $M = 0.6256$ and compute the value of the sign test statistic for each sample generated, i.e., the number of X_i in that sample for which $X_i - M_0 = X_i - 0.5 > 0$. Then we note whether or not this count value is in the rejection region $K \geqslant 10$. Then we count the number of times we found the count value in the rejection region among the 1000 random samples generated. This count divided by 1000 is the simulated power at the point $\theta = 0.55$ (which corresponds to $M_1 = 0.6256$) in the case $N = 13$, $M_0 = 0.50$, $\sigma = 1$, and $\alpha = 0.0461$. Using a MINITAB Macro program, this value was found as 0.10. Note that from Table 4.1, the normal approximation to the power in this case is 0.0918. The program code is shown below for this situation:

```
macro sign
# simulates power of sign test with N=13, alpha =0.0461
sign
mcolumn c1 c2 c10 pow1
mconstant mu k1 k3 k5 cow theta
let k5=1
let mu=0.50
let theta=0.0
mlabel 4
let cow=1
let pow1(k5)=0
mlabel 1
random 13 c1;
normal mu 1.
let c1=c1-.50
# calculate Sign test statistic
let c2 = (c1 gt 0)
let k1 = sum(c2)
# k1 = value of Sign stat
let k3 = 10
if k1 ge 10
let pow1(k5)=pow1(k5)+1
 else
 let pow1(k5)=pow1(k5)+0
 endif
 let cow=cow+1
```


Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

```
if cow gt 1000
go to 2
else
go to 1
endif
mlabel 2
let pow1(k5)=pow1(k5)/1000
print k5 pow1(k5)
let k5=k5+1
let mu=mu+.05
if mu gt 0.9
go to 3
else
go to 4
endif
mlabel 3
set c10
.5(.05).9
end
print c10 pow1
plot pow1*c10;
  connect;
  color 1 2;
axis 2;
  label 'simulated power'.
endmacro
```

To run such a program, type the statements into a plain text file, using a text editor (not a word processor) and save it with a .mac extension to a floppy disk, say, in drive a. Suppose the name of the file is sign.mac. Then in MINITAB, go to edit, then to command line editor and then type % a:\sign.mac and click on submit. The program will print the simulated power values as well as a power curve. Output from such a simulation is shown later in Section 5.7 as Figure 7.1.

SAMPLE SIZE DETERMINATION

In order to make an inference regarding the population median using the sign test, we need to have a random sample of observations. If we are allowed to choose the sample size, we might want to determine the value of N such that the test has size α and power $1-\beta$, given the null and the alternative hypotheses and other necessary assumptions. For example, for the sign test against the one-sided upper-tailed alternative H_1: $M > M_0$, we need to find N such that

$$\sum_{i=k_\alpha}^{N}\binom{N}{i}(0.5)^N \leqslant \alpha \qquad \text{and} \qquad \sum_{i=k_\alpha}^{N}\binom{N}{i}\theta^i(1-\theta)^{N-i} \geqslant 1-\beta$$

where α, $1-\beta$ and $\theta = P(X > M_0|H_1)$ are all specified. Note also that the size and the power requirements have been modified to state "at most" α and "at least" $1-\beta$, in order to reflect the discreteness of the binomial distribution. Tables are available to aid in solving these equations; see for example, Cohen (1972). We illustrate the process using the normal approximation to the power because the necessary equations are much easier to solve.

Under the normal approximation, the power of a size α sign test with H_1: $M > M_0$ is given in (4.8). Thus we require that $1-\Phi\left[N(0.5-\theta) + 0.5\sqrt{N}z_\alpha/\sqrt{N\theta(1-\theta)}\right] = 1-\beta$ or, solving for N, we get

$$N = \left[\frac{\sqrt{\theta(1-\theta)}\Phi^{-1}(\beta) - 0.5z_\alpha}{0.5-\theta}\right]^2 = \left[\frac{\sqrt{\theta(1-\theta)}z_\beta + 0.5z_\alpha}{0.5-\theta}\right]^2 \tag{4.9}$$

which should be rounded up to the next integer. The approximate sample size formula for the one-sided lower-tailed alternative H_1: $M < M_0$ is the same except that here $\theta = P(X > M_0|H_1) < 0.5$. A sample size formula for the two-sided alternative is the same as (4.9) with α replaced by $\alpha/2$. The derivation is left as an exercise for the reader.

For example, suppose $\theta = 0.2$. If we set $\alpha = 0.05$ and $\beta = 0.90$, then $z_\alpha = 1.645$ and $z_\beta = 1.282$. Then (4.9) yields $\sqrt{N} = 4.45$ and $N = 19.8$. Thus we need at least 20 observations to meet the specifications.

CONFIDENCE INTERVAL FOR THE MEDIAN

A two-sided confidence-interval estimate for an unknown population median can be obtained from the acceptance region of the sign test against the two-sided alternative. The acceptance region for a two-sided test of H_0: $M = M_0$, using (4.3), is

$$k'_{\alpha/2} + 1 \leqslant K \leqslant k_{\alpha/2} - 1 \tag{4.10}$$

where K is the number of positive differences among $X_i - M$, $i = 1, 2, \ldots, N$ and $k'_{\alpha/2}$ and $k_{\alpha/2}$ are integers such that

$$P(k'_{\alpha/2} + 1 \leqslant K \leqslant k_{\alpha/2} - 1) \geqslant 1-\alpha$$

As we found for the quantile test, the equal-tailed confidence interval endpoints for the unknown population median are the order statistics

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

$X_{(r)}$ and $X_{(s)}$ where r and s are the largest and smallest integers respectively, such that

$$\sum_{i=0}^{r-1} \binom{N}{i} (0.5)^N \leqslant \frac{\alpha}{2} \qquad \text{and} \qquad \sum_{i=s}^{N} \binom{N}{i} (0.5)^N \leqslant \frac{\alpha}{2} \tag{4.11}$$

We note that $r - 1$ and s are easily found from Table G in the columns labeled Left tail and Right tail, respectively.

For larger sample sizes,

$$r = k'_{\alpha/2} + 1 = 0.5 + 0.5N - 0.5\sqrt{N} z_{\alpha/2} \tag{4.12}$$

and

$$s = k_{\alpha/2} = 0.5 + 0.5N + 0.5\sqrt{N} z_{\alpha/2} \tag{4.13}$$

We round down for r and round up for s for a conservative solution.

In order to contrast the exact and approximate confidence interval endpoints suppose $N = 15$ and $1 - \alpha = \gamma = 0.95$. Then, using Table G with $\theta = 0.5$, $r = 4$ for significance level 0.0176 so that the exact endpoints of the 95% confidence interval are $X_{(4)}$ and $X_{(12)}$ with exact confidence level $\gamma = 0.9648$. For the approximate confidence interval $r = 0.5 + 7.5 - 0.5\sqrt{15}(1.65) = 4.21$ which we round down. So the confidence interval based on the normal approximation is also given by $(X_{(4)}, X_{(12)})$ with exact confidence level $\gamma = 0.9648$.

PROBLEM OF ZEROS

Zeros do not present a problem in finding a confidence interval estimate of the median using this procedure. As a result, the sample size N is not reduced for zeros and zeros are counted as many times as they occur in determining confidence-interval endpoints. If the real interest is in hypothesis testing and there are many zeros, the power of the test will be greater if the test is carried out using a confidence-interval approach.

PAIRED-SAMPLE PROCEDURES

The one-sample sign-test procedures for hypothesis testing and confidence interval estimation of M are equally applicable to paired-sample data. For a random sample of N pairs $(X_1,Y_1), \ldots, (X_N,Y_N)$, the N differences $D_i = X_i - Y_i$ are formed. If the population of differences is assumed continuous at its median M_D so that $P(D = M_D) = 0$, and θ is defined as $\theta = P(D > M_D)$, the same procedures are clearly valid here with X_i replaced everywhere by D_i.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

It should be emphasized that this is a test for the median difference M_D, which is not necessarily the same as the difference of the two medians M_X and M_Y. The following simple example will serve to illustrate this often misunderstood fact. Let X and Y have the joint distribution

$$f_{X,Y}(x,y) = \begin{cases} 1/2 & \text{for } y-1 \leqslant x \leqslant y, -1 \leqslant y \leqslant 1 \\ & \quad \text{or } y+1 \leqslant x \leqslant 1, -1 \leqslant y \leqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

Then X and Y are uniformly distributed over the shaded region in Figure 4.2. It can be seen that the marginal distributions of X and Y are identical, both being uniform on the interval $(-1,1)$, so that $M_X = M_Y = 0$. It is clear that where X and Y have opposite signs, in quadrants II and IV,

$$P(X < Y) = P(X > Y)$$

while in quadrants I and III, $X < Y$ always. For all pairs, then, we have $P(X < Y) = 3/4$, which implies that the median of the population of differences is smaller than zero. It will be left as an exercise for the reader to show that the cdf of the difference random variable $D = X - Y$ is

$$F_D(d) = \begin{cases} 0 & \text{for } d \leqslant -1 \\ (d+1)(d+3)/4 & \text{for } -1 < d \leqslant 0 \\ 3/4 & \text{for } 0 < d \leqslant 1 \\ d(4-d)/4 & \text{for } 1 < d \leqslant 2 \\ 1 & \text{for } d > 2 \end{cases} \tag{4.14}$$

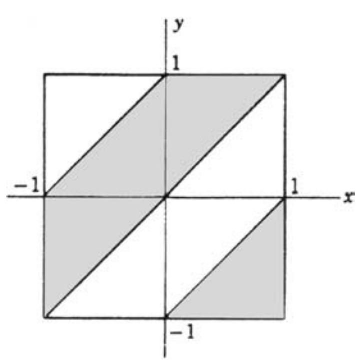


Fig. 4.2 Region of integration is the shaded area.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The median difference is that value M_D, of the distribution of D, such that $F_D(M_D) = 1/2$. The reader can verify that this yields $M_D = -2 + \sqrt{3}$.

In general, then, it is not true that $M_D = M_X - M_Y$. On the other hand, it is true that a mean of differences equals the difference of means. Since the mean and median coincide for symmetric distributions, if the X and Y populations are both symmetric and $M_X = M_Y$, and if the difference population is also symmetric,[1] then $M_D = M_X - M_Y$ and $M_X = M_Y$ is a necessary and sufficient condition for $M_D = 0$. Note that for the case where X and Y are each normally distributed, the difference of their medians (or means) is equal to the median (or mean) of their difference $X - Y$, since $X - Y$ is also normally distributed with median (or mean) equal to the difference of the respective medians (or means).

Earlier discussions of power and sample size also apply to the paired-sample data problems.

APPLICATIONS

We note that the sign test is a special case of the quantile test with $p = 0.5$, since the quantile specified is the population median. This test is easier to apply than the general quantile test because the binomial distribution for $\theta = 0.5$ is symmetric for any N. We write the null hypothesis here as H_0: $M = M_0$. The appropriate rejection regions in terms of K, the number of plus signs among $X_1 - M_0, X_2 - M_0, \ldots, X_N - M_0$, and corresponding exact P values, are summarized as follows:

Alternative	*Rejection region*	*Exact P value*
$M > M_0$	$K \geqslant k_\alpha$	$\sum_{i=K_O}^{N} \binom{N}{i} (0.5)^N$
$M < M_0$	$K \leqslant k'_\alpha$	$\sum_{i=0}^{K_O} \binom{N}{i} (0.5)^N$
$M \neq M_0$	$K \leqslant k'_{\alpha/2}$ or $K \geqslant k_{\alpha/2}$	2(smaller of the one-tailed P values)

Table C with $\theta = 0.5$ and n (representing N) can be used to determine the critical values. Table G is simpler to use because it gives both left-tail and right-tail binomial probabilities for $N \leqslant 20$ when $\theta = 0.5$.

[1]The difference population is symmetric if X and Y are symmetric and independent or if $f_{X,Y}(x, y) = f_{X,Y}(-x, -y)$.

For large sample sizes, the appropriate rejection regions and the P values, based on the normal approximation to the binomial distribution with a continuity correction, are as follows:

Alternative	*Rejection region*	*Approximate P value*
$M > M_0$	$K \geqslant 0.5N + 0.5 + 0.5z_\alpha\sqrt{N}$	$1 - \Phi\left(\dfrac{K_O - 0.5N - 0.5}{0.5\sqrt{N}}\right)$
$M < M_0$	$K \geqslant 0.5N - 0.5 - 0.5z_\alpha\sqrt{N}$	$\Phi\left(\dfrac{K_O - 0.5N + 0.5}{0.5\sqrt{N}}\right)$
$M \neq M_0$	Both above with $z_{\alpha/2}$	2(smaller of the one-tailed P values)

If any zeros are present, we will ignore them and reduce N accordingly. As we have seen, a prespecified significance level α often cannot be achieved with nonparametric statistical inference because most of the applicable sampling distributions are discrete. This problem is avoided if we determine the P value of a test result and use that to make our decision.

For a two-sided alternative, the common procedure is to define the P value as twice the smaller of the two one-sided P values, as described in the case for general quantiles. The "doubling" is particularly meaningful when the null distribution of the test statistic is symmetric, as is the case here. For example, suppose that we observe four plus signs among $N = 12$ nonzero sample differences. Table G shows that the left-tail P value is 0.1938; since there is no entry in the right-tail column, we know that the right-tail P-value exceeds 0.5. Thus the two-sided P value is 2 times 0.1938, or 0.3876.

Another way of looking at this is as follows. Under the null hypothesis the binomial distribution is symmetric about the expected value of K, which here is $N(0.5) = 6$. Thus, for any value of K less than 6, the upper-tail probability will be greater than 0.5 and the lower-tail probability less than 0.5. Conversely, for any value of K greater than 6, the upper-tail probability is less than 0.5 and the lower-tail probability is greater than 0.5. Also, by symmetry, the probability of say 4 or less is the same as the probability of 8 or more. Thus, to calculate the P value for the two-sided alternative, the convention is to take the smaller of the two one-tailed P values and double it. If instead we used the larger of the P values and doubled that, the final P value could possibly be more than 1.0, which is not acceptable. Note also that when the observed value of K is exactly equal to 6, the two-sided P value will be taken to be equal to 1.0.

In our example, the observed value 4 for $N=12$ is less than 6, so the smaller one-tailed P value is in the lower tail and is equal to 0.1938 and this leads to a two-sided P value of 0.3876 as found earlier. If we have a prespecified α, and wish to reach a decision, we should reject H_0 whenever the P value is less than or equal to α and accept H_0 otherwise.

The exact distribution-free confidence interval for the median can be found from Table C but is particularly easy to find using Table G. The choice of exact confidence levels is limited to $1-2P$, where P is a tail probability in Table G for the appropriate value of N. From (4.10), the lower confidence limit is the $(k'_{\alpha/2}+1)$th $= r$th-order statistic in the sample, where $k'_{\alpha/2}$ is the left-tail critical value of the sign test statistic K from Table G, for the given α and N such that the P figure is less than or equal to $\alpha/2$. But since the critical values are all of the nonnegative integers, $k'_{\alpha/2}+1$ is simply the rank of $k'_{\alpha/2}$ among the entries in Table G for that N. The calculation of this rank will become clearer after we do Example 4.1.

For consistency with the results given later for confidence interval endpoints based on other nonparametric test procedures, we note that r is the rank of the left-tail entry in Table G for this N, and we denote this rank by u. Further, by symmetry, we have $X_{(s)} = X_{(N-r+1)}$. The confidence interval endpoints are the uth from the smallest and the uth from the largest order statistics, where u is the rank of the left-tail critical value of K from Table G that corresponds to $P \leqslant \alpha/2$. The corresponding exact confidence coefficient is then $\gamma = 1-2P$. For sample sizes outside the range of Table G we have

$$u = 0.5 + 0.5N - 0.5\sqrt{N}z_{\alpha/2} \tag{4.15}$$

from (4.4), and we always round the result of (4.15) downward.

For example, for a confidence level of 0.95 with $N=15$, $\alpha/2 = 0.025$, the P figure from Table G closest to 0.025 but not exceeding it is 0.0176. The corresponding left-tail critical value is 3, which has a rank of 4 among the left-tail critical values for this N. Thus $u=4$ and the 95% confidence interval for the median is given by the interval $(X_{(4)}, X_{(12)})$. The exact confidence level for this distribution-free interval is $1-2P = 1-2(0.0176) = 0.9648$.

Note that unlike in the case of testing hypotheses, if zeros occur in the data, they are counted as many times as they appear for determination of the confidence interval endpoints.

Example 4.1 Suppose that each of 13 randomly chosen female registered voters was asked to indicate if she was going to vote for

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

candidate A or candidate B in an upcoming election. The results show that 9 of the subjects preferred A. Is this sufficient evidence to conclude that candidate A is preferred to B by female voters?

Solution With this kind of data, the sign test is one of the few statistical tests that is valid and can be applied. Let θ be the true probability that candidate A is preferred over candidate B. The null hypothesis is that the two candidates are equally preferred, that is, H_0: $\theta = 0.5$ and the one-sided upper-tailed alternative is that A is preferred over B, that is H_1: $\theta > 0.5$. The sign test can be applied here and the value of the test statistic is $K = 9$. Using Table G with $N = 13$, the exact P value in the right-tail is found to be 0.1338; therefore this is not sufficient evidence to conclude that the female voters prefer A over B, at a commonly used significant level such as 0.05.

Example 4.2 Some researchers claim that susceptibility to hypnosis can be acquired or improved through training. To investigate this claim six subjects were rated on a scale of 1 to 20 according to their initial susceptibility to hypnosis and then given 4 weeks of training. Each subject was rated again after the training period. In the ratings below, higher numbers represent greater susceptibility to hypnosis. Do these data support the claim?

Subject	*Before*	*After*
1	10	18
2	16	19
3	7	11
4	4	3
5	7	5
6	2	3

Solution The null hypothesis is H_0: $M_D = 0$ and the appropriate alternative is H_1: $M_D > 0$ where M_D is the median of the differences, after training minus before training. The number of positive differences is $K_O = 4$ and the right-tail P value for $N = 6, K_O = 4$ from Table G is 0.3438. Hence the data do not support the claim at any level smaller than 0.3438 which implies that 4 is not an extreme value of K under H_0; rejection of the null hypothesis is not warranted. Also, from Table G, at $\alpha = 0.05$, the rejection region is $K \geqslant 6$, with exact size 0.0156. Since the observed value of K equals 4, we again fail to reject H_0.

The following computer printouts illustrate the solution to Example 4.2 based on the STATXACT, MINITAB, and SAS packages. The

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

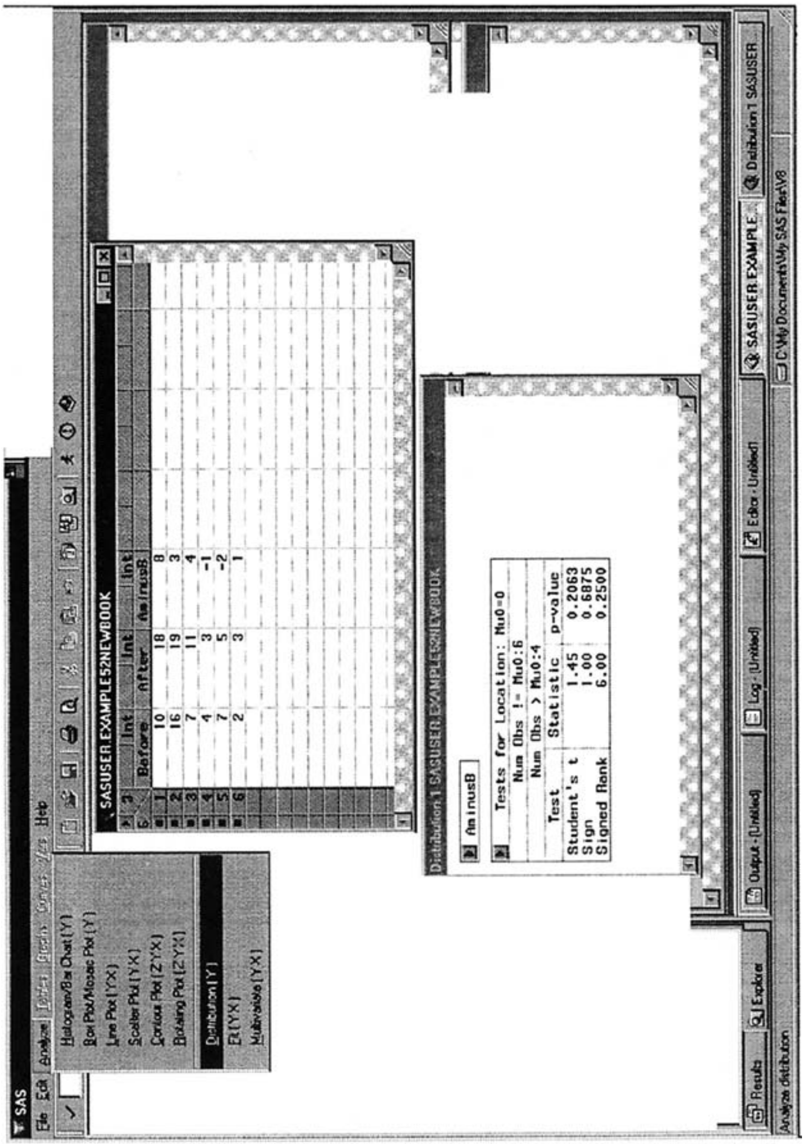
STATXACT solution agrees with ours for the exact one-sided P value. Their asymptotic P value (0.2071) is based on the normal approximation without the continuity correction. The MINITAB solution agrees exactly with ours. The SAS solution gives only the two-tailed P values. The exact sign test result in 2 times ours; they also give P values based on Student's t test and the signed-rank test discussed later in this chapter.

```
                  **********************************
                  STATXACT SOLUTION TO EXAMPLE 4.2
                  **********************************

SIGN TEST

Summary of Exact distribution of SIGN statistic:
Min         Max         Mean        Std-dev      Observed Standardized
0.0000      6.000       3.000       1.225        2.000       -0.8165

Asymptotic Inference:
One-sided p-value: Pr { Test Statistic .LE. Observed }  =    0.2071
Two-sided p-value: 2 * One-sided                        =    0.4142

Exact Inference:
One-sided p-value: Pr { Test Statistic .LE. Observed }  =    0.3438
Pr { Test Statistic .EQ. Observed }  =    0.2344
Two-sided p-value: 2*One-Sided                          =    0.6875


                  **********************************
                  MINITAB SOLUTION TO EXAMPLE 4.2
                  **********************************

Sign Test for Median: Af-Be

Sign test of median = 0.00000 versus  >  0.00000

               N  Below  Equal  Above        P     Median
Af-Be          6      2      0      4   0.3438      2.000
```

Now suppose we wanted to know, before the investigation started, how many subjects should be included in the study when we plan to use the sign test for the median difference at a level of significance $\alpha = 0.05$, and we want to detect $P(D > 0) = 0.6$ with a power 0.85. Note that $P(D > 0) = 0.6$ means that the median difference, M_D, is greater than 0, the hypothesized value, and thus the test should have an upper-tailed alternative. With $\theta = 0.6, z_{0.05} = 1.645$, and $z_{0.15} = 1.0365$, Eq. (4.9) gives $N = 176.96$ which we round up to 177.

The MINITAB solution to this example is shown below. It also uses the normal approximation and the result 177 agrees with ours.

****
SAS SOLUTION TO EXAMPLE 4.2
****



MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

The solution also shows $N = 222$ observations will be required for a two-tailed test. The reader can verify this. The solution is labeled "Test for One Proportion" instead of "Sign Test" because it is applicable for a test for a quantile of any order p (as in Section 5.3).

```
******************************************
MINITAB SOLUTION TO POWER AND SAMPLE SIZE
******************************************

Power and Sample Size

Test for One Proportion

Testing proportion = 0.5 (versus > 0.5)
Calculating power for proportion = 0.6
Alpha = 0.05  Difference = 0.1

Sample   Target    Actual
 Size    Power     Power
 177     0.8500    0.8501
```

```
Power and Sample Size

Test for One Proportion

Testing proportion = 0.5 (versus not = 0.5)
Calculating power for proportion = 0.6
Alpha = 0.05  Difference = 0.1

Sample   Target    Actual
 Size    Power     Power
 222     0.8500    0.8511
```

Example 4.3 Nine pharmaceutical laboratories cooperated in a study to determine the median effective dose level of a certain drug. Each laboratory carried out experiments and reported its effective dose. For the results 0.41, 0.52, 0.91, 0.45, 1.06, 0.82, 0.78, 0.68, 0.75, estimate the interval of median effective dose with a confidence level 0.95.

Solution We go to Table G with $N = 9$ and find $P = 0.0195$ is the largest entry that does not exceed 0.025, and this entry has rank $u = 2$. Hence the second smallest and second largest (or the $9 - 2 + 1 = 8$th smallest) order statistics of the sample data, namely $X_{(2)}$ and $X_{(8)}$, provide the two endpoints as $0.45 < M < 0.91$ with exact confidence coefficient $1 - 2(0.0195) = 0.961$. The MINITAB solution

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

shown gives the two confidence intervals with the exact confidence coefficient on each side of 0.95, as well as an exact 95% interval, based on an interpolation scheme between the two sets of endpoints, lower and upper, respectively. This latter interval is indicated by NLI on the output. The interpolation scheme is a nonlinear one due to Hettmansperger and Sheather (1986).

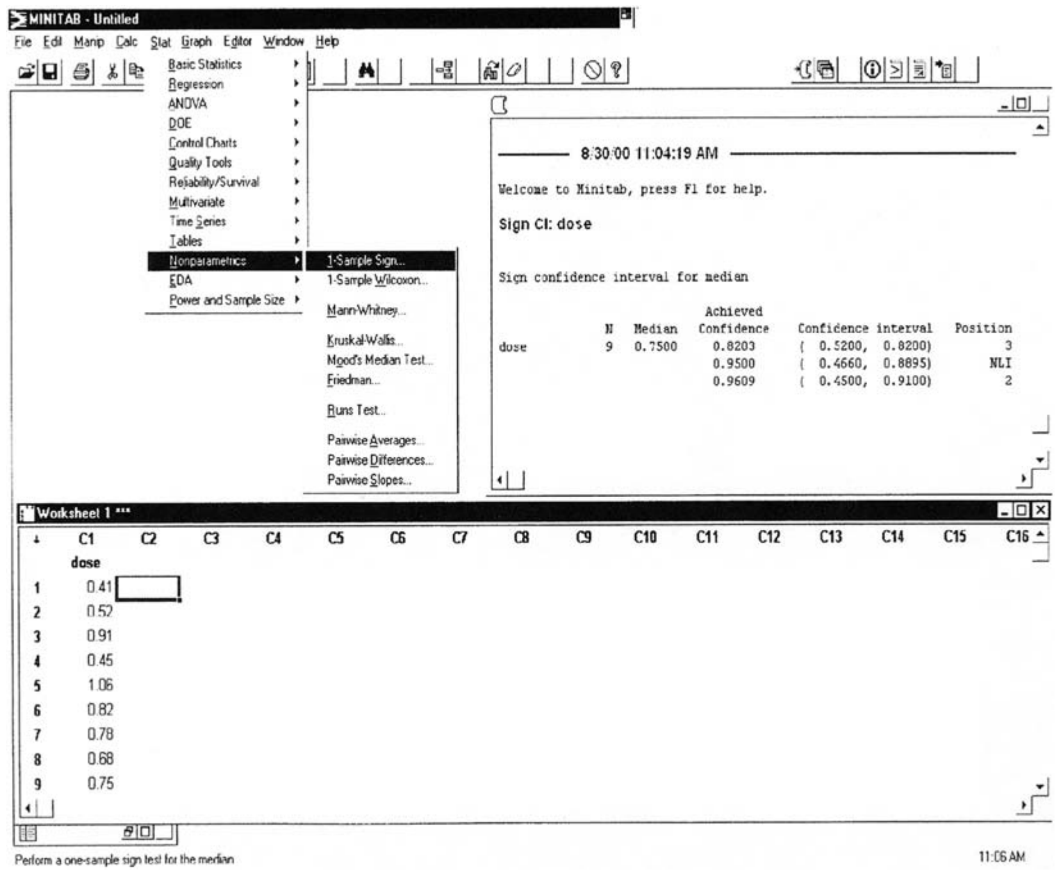
5.5 RANK-ORDER STATISTICS

The other one-sample procedure to be covered in this chapter in the Wilcoxon signed-rank test. This test is based on a special case of what are called rank-order statistics. The *rank-order statistics* for a random sample are any set of constants which indicate the order of the observations. The actual magnitude of any observation is used only in determining its relative position in the sample array and is thereafter ignored in any analysis based on rank-order statistics. Thus any statistical procedures based on rank-order statistics depend only on the relative magnitudes of the observations. If the jth element X_j is the ith smallest in the sample, the jth rank-order statistics must be the ith smallest rank-order statistic. Rank-order statistics might then be defined as the set of numbers which results when each original observation is replaced by the value of some order-preserving function. Suppose we have a random sample of N observations $X_1, X_2, \ldots, X_N$. Let the rank-order statistics be denoted by $r(X_1), r(X_2), \ldots, r(X_N)$ where r is any function such that $r(X_i) \leqslant r(X_j)$ whenever $X_i \leqslant X_j$. As with order statistics, rank-order statistics are invariant under monotone transformations, i.e., if $r(X_i) \leqslant r(X_j)$, then $r[F(X_i)] \leqslant r[F(X_j)]$, in addition to $F[r(X_i)] \leqslant F[r(X_j)]$, where F is any nondecreasing function.

For any set of N different sample observations, the simplest set of numbers to use to indicate relative positions is the first N positive integers. In order to eliminate the possibility of confusion and to simplify and unify the theory of rank-order statistics, we shall assume here that unless explicitly stated otherwise, the rank-order statistics are always a permutation of the first N integers. The ith rank-order statistic $r(X_i)$ then is called the rank of the ith observation in the original unordered sample. The value it assumes, $r(x_i)$, is the number of observations x_j, $j = 1, 2, \ldots, N$, such that $x_j \leqslant x_i$. For example, the rank of the ith-order statistic is equal to i, or $r(x_{(i)}) = i$. A functional definition of the rank of any x_i in a set of N different observations is provided by

$$r(X_i) = \sum_{j=1}^{N} S(x_i - x_j) = 1 \sum_{j \neq 1} S(x_i - x_j) \tag{5.1}$$

****
MINITAB SOLUTION TO EXAMPLE 4.3
****



Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where

$$S(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u \geqslant 0 \end{cases} \tag{5.2}$$

The random variable $r(X_i)$ is discrete and for a random sample from a continuous population it follows the discrete uniform distribution, or

$$P[r(X_i) = j] = 1/N \qquad \text{for } j = 1, 2, \ldots, N$$

Although admittedly the terminology may seen confusing at the outset, a function of the rank-order statistics will be a called a *rank statistic*. Rank statistics are particularly useful in nonparametric inference since they are usually distribution free. The methods are applicable to a wide variety of hypothesis-testing situations depending on the particular function used. The procedures are generally simple and quick to apply. Since rank statistics are functions only of the ranks of the observations, only this information is needed in the sample data. Actual measurements are often difficult, expensive, or even impossible to obtain. When actual measurement are not available for some reason but relative positions can be determined, rank-order statistics make use of all of the information available. However, when the fundamental data consist of variate values and these actual magnitudes are ignored after obtaining the rank-order statistics, we may be concerned about the loss of efficiency that may ensue. One approach to a judgment concerning the potential loss of efficiency is to determine the correlation between the variate values and their assigned ranks. If the correlation is high, we would feel intuitively more justified in the replacement of actual values by ranks for the purpose of analysis. The hope is that inference procedures based on ranks alone will lead to conclusions which seldom differ from a corresponding inference based on actual variate values.

The ordinary product-moment correlation coefficient between two random variables X and Y is

$$\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$$

Assume that for a continuous population denoted by a cdf F_X (pdf f_X) we would like to determine the correlation between the random variable X and its rank $r(X)$. Theoretically, a random variable from an infinite population cannot have a rank, since values on a continuous scale cannot be ordered. But an observation X_i, of a random sample of size N from this population, does have a rank $r(X_i)$ as defined in (5.1).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The distribution of X_i is the same as the distribution of X and the $r(X_i)$ are identically distributed though not independent. Therefore, it is reasonable to define the population correlation coefficient between ranks and variate values as the correlation between X_i and $Y_i = r(X_i)$, or

$$\rho[X, r(X)] = \frac{E(X_i Y_i) - E(X_i)E(Y_i)}{\sigma_X \sigma_{Y_i}} \tag{5.3}$$

The marginal distribution of Y_i for any i is the discrete uniform, so that

$$f_{Y_i}(j) = \frac{1}{N} \qquad \text{for } j = 1, 2, \ldots, N \tag{5.4}$$

with moments

$$E(Y_i) = \sum_{j=1}^{N} \frac{j}{N} = \frac{N+1}{2} \tag{5.5}$$

$$E(Y_i^2) = \sum_{j=1}^{N} \frac{j^2}{N} = \frac{(N+1)(2N+1)}{6}$$

$$\operatorname{var}(Y_i) = \frac{(N+1)(2n+1)}{6} - \frac{(N+1)^2}{4} = \frac{N^2-1}{12} \tag{5.6}$$

The joint pdf of X_i and its rank Y_i is

$$f_{X_i,Y_i}(x,j) = f_{X_i|Y_i=j}(x|j) f_{Y_i}(j) = \frac{f_{X_{(j)}}(x)}{N} \qquad \text{for } j = 1, 2, \ldots, N$$

where $X_{(j)}$ denotes the jth-order statistic of a random sample of size N from the cdf F_X. From this expression we can write

$$E(X_i Y_i) = \frac{1}{N} \int_{-\infty}^{\infty} \sum_{j=1}^{N} j x f_{X_{(j)}}(x)\, dx = \sum_{j=1}^{N} \frac{jE(X_{(j)})}{N} \tag{5.7}$$

Substituting the results (5.5), (5.6), and (5.7) back into (5.3), we obtain

$$\rho[X, r(X)] = \left(\frac{12}{N^2-1}\right)^{1/2} \frac{\sum_{j=1}^{N} jE(X_{(j)}) - [N(N+1)/2]E(X)}{N\sigma_X} \tag{5.8}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Since the result here is independent of i, our definition in (5.3) may be considered a true correlation. The same result is obtained if the covariance between X and $r(X)$ is defined as the limit as M approaches infinity of the average of the M correlations that can be calculated between sample values and their ranks when M samples of size N are drawn from this population. This method will be left as an exercise for the reader.

The expression given in (5.8) can be written in another useful form. If the variate values X are drawn from a continuous population with distribution F_X, the following sum can be evaluated:

$$\sum_{i=1}^{N} iE(X_{(i)}) = \sum_{i=1}^{N} \frac{iN!}{(i-1)!(N-i)!}$$
$$\times \int_{-\infty}^{\infty} x[F_X(x)]^{i-1}[1-F_X(x)]^{N-i} f_X(x)dx$$
$$= \sum_{j=0}^{N-1} \frac{(j+1)N!}{j!(N-j-1)!} \int_{-\infty}^{\infty} x[F_X(x)]^{j}[1-F_X(x)]^{N-j-1} f_X(x)dx$$
$$= \sum_{j=1}^{N-1} \frac{N!}{(j-1)!(N-j-1)!}$$
$$\times \int_{-\infty}^{\infty} x[F_X(x)]^{j}[1-F_X(x)]^{N-j-1} f_X(x)dx$$
$$+ \sum_{j=0}^{N-1} \frac{N!}{j!(N-j-1)!} \int_{-\infty}^{\infty} x[F_X(x)]^{j}[1-F_X(x)]^{N-j-1} f_X(x)dx$$
$$= N(N-1) \int_{-\infty}^{\infty} xF_X(x)$$
$$\times \sum_{j=1}^{N-1} \binom{N-2}{j-1} [F_X(x)]^{j-1}[1-F_X(x)]^{N-j-1} f_X(x)\,dx$$
$$+ N \int_{-\infty}^{\infty} x \sum_{j=0}^{N-1} \binom{N-1}{j} [F_X(x)]^{j}[1-F_X(x)]^{N-j-1} f_X(x)\,dx$$
$$= N(N-1) \int_{-\infty}^{\infty} xF_X(x) f_X(x)dx + N \int_{-\infty}^{\infty} xf_X(x)\,dx$$
$$= N(N-1)E[XF_X(x)] + NE(X) \tag{5.9}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

If this quantity is now substituted in (5.8), the result is

$$\begin{aligned}\rho[X,r(X)] &= \left(\frac{12}{N^2-1}\right)^{1/2}\frac{1}{\sigma_X}\left\{(N-1)E[XF_X(X)]+E(X)-\frac{N+1}{2}E(X)\right\}\\ &= \left(\frac{12}{N^2-1}\right)^{1/2}\frac{1}{\sigma_X}\left\{(N-1)E[XF_X(X)]-\frac{N-1}{2}E(X)\right\}\\ &= \left[\frac{12(N-1)}{N+1}\right]^{1/2}\frac{1}{\sigma_X}\left\{E[XF_X(X)]-\frac{1}{2}E(X)\right\} \qquad (5.10)\end{aligned}$$

and

$$\lim_{N\to\infty}\rho[X,r(X)] = \frac{2\sqrt{3}}{\sigma_X}\left\{E[XF_X(X)]-\frac{1}{2}E(X)\right\} \qquad (5.11)$$

Some particular evaluations of (5.11) are given in Stuart (1954).

5.6 TREATMENT OF TIES IN RANK TESTS

In applying tests based on rank-order statistics, we usually assume that the population from which the sample was drawn is continuous. When this assumption is made, the probability of any two observations having identical magnitudes is equal to zero. The set of ranks as defined in (5.1) then will be N different integers. The exact properties of most rank statistics depend on this assumption. Two or more observations with the same magnitude are said to be *tied*. We may say only that *theoretically* no problem is presented by tied observations. However, in practice ties can certainly occur, either because the population is actually discrete or because of practical limitations on the precision of measurement. Some of the conventional approaches to dealing with ties in assigning ranks will be discussed generally in this section, so that the problem can be ignored in presenting the theory of some specific rank tests later.

In a set of N observations which are *not* all different, arrangement in order of magnitude produces a set of r groups of different numbers, the ith different value occurring with frequency t_i, where $\sum t_i = N$. Any group of numbers with $t_i \geqslant 2$ comprises a set of tied observations. The ranks are no longer well defined, and for any set of fixed ranks of N untied observations there are $\prod t_i!$ possible assignments of ranks to the entire sample with ties, each assignment leading to its own value for a rank test statistic, although that value may be the same as for some other assignment. If a rank test is to be performed using a sample containing tied observations, we must have

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

either a unique method of assigning ranks for ties so that the test statistic can be computed in the usual way or a method of combining the many possible values of the rank test statistic to reach one decision. Several acceptable methods will be discussed briefly.

RANDOMIZATION

In the method of randomization, one of the $\prod t_i!$ possible assignments of ranks is selected by some random procedure. For example, in the set of observations

$$3.0, 4.1, 4.1, 5.2, 6.3, 6.3, 6.3, 9$$

there are 2!(3!) or 12 possible assignments of the integer ranks 1 to 8 which this sample could represent. One of these 12 assignments is selected by a supplementary random experiment and used as the unique assignment of ranks. Using this method, some theoretical properties of the rank statistic are preserved, since each assignment occurs with equal probability. In particular, the null probability distribution of the rank-order statistic, and therefore of the rank statistic, is unchanged, so that the test can be performed in the usual way. However, an additional element of chance is artificially imposed, affecting the probability distribution under alternatives.

MIDRANKS

The midrank method assigns to each member of a group of tied observations the simple average of the ranks they would have if distinguishable. Using this approach, tied observations are given tied ranks. The midrank method is perhaps the most frequently used, as it has much appeal experimentally. However, the null distribution of ranks is affected. Obviously, the mean rank is unchanged, but the variance of the ranks would be reduced. When the midrank method is used, for some tests a correction for ties can be incorporated into the test statistic. We discuss these corrections when we present the respective tests.

AVERAGE STATISTIC

If one does not wish to choose a particular set of ranks as in the previous two methods, one may instead calculate the value of the test statistic for all the $\prod t_i!$ assignments and use their simple average as the single sample value. Again, the test statistic would have the same mean but smaller variance.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

AVERAGE PROBABILITY

Instead of averaging the test statistic for each possible assignment of ranks, one could find the probability of each resulting value of the test statistic and use the simple average of these probabilities for the overall probability. This requires availability of tables of the exact null probability distribution of the test statistic rather than simply a table of critical values.

LEAST FAVORABLE STATISTIC

Having found all possible values of the test statistic, one might choose as a single value that one which minimizes the probability of rejection. This procedure leads to the most conservative test, i.e., the lowest probability of committing a type I error.

RANGE OF PROBABILITY

Alternatively, one could compute two values of the test statistic: the one least favorable to rejection and the one most favorable. However, unless both fall inside or both fall outside the rejection region, this method does not lead to a decision.

OMISSION OF TIED OBSERVATIONS

The final and most obvious possibility is to discard all tied observations and reduce the sample size accordingly. This method certainly leads to a loss of information, but if the number of observations to be omitted is small relative to the sample size, the loss may be minimal. This procedure generally introduces bias toward rejection of the null hypothesis.

The reader is referred to Savage's *Bibliography* (1962) for discussions of treatment of ties in relation to particular nonparametric rank test statistics. Pratt and Gibbons (1981) also give detailed discussions and many references. Randles (2001) gives a different approach to dealing with ties.

5.7 THE WILCOXON SIGNED-RANK TEST AND CONFIDENCE INTERVAL

Since the one-sample sign test in Section 5.4 utilizes only the signs of the differences between each observation and the hypothesized median M_0, the magnitudes of these observations relative to M_0 are ignored. Assuming that such information is available, a test statistic which takes into account these individual relative magnitudes might

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

be expected to give better performance. If we are willing to make the assumption that the parent population is symmetric, the Wilcoxon signed-rank test statistic provides an alternative test of location which is affected by both the magnitudes and signs of these differences. The rationale and properties of this test will be discussed in this section.

As with the one-sample situation of Section 5.4, we have a random sample of N observations $X_1, X_2, \ldots, X_N$ from a continuous cdf F with median M, but now we assume that F is symmetric about M. Under the null hypothesis

$$H_0: M = M_0$$

the differences $D_i = X_i - M_0$ are symmetrically distributed about zero, so that positive and negative differences of equal absolute magnitude have the same probability of occurrence; i.e., for any $c > 0$,

$$F_D(-c) = P(D_i \leqslant -c) = P(D_i \geqslant c) = 1 - P(D_i \leqslant c) = 1 - F_D(c)$$

With the assumption of a continuous population, we need not be concerned theoretically with zero or tied absolute differences $|D_i|$. Suppose we order these absolute differences $|D_1|, |D_2|, \ldots, |D_N|$ from smallest to largest and assign them ranks $1, 2, \ldots, N$, keeping track of the original signs of the differences D_i. If M_0 is the true median of the symmetrical population, the expected value of the sum of the ranks of the positive differences T^+ is equal to the expected value of the sum of the ranks of the negative differences T^-. Since the sum of all the ranks is a constant, that is, $T^+ + T^- = \sum_{i=1}^{N} i = N(N+1)/2$, test statistics based on T^+ only, T^- only, or $T^+ - T^-$ are linearly related and therefore equivalent criteria. In contrast to the ordinary one-sample sign test, the value of T^+, say, is influenced not only by the number of positive differences but also by their relative magnitudes. When the symmetry assumption can be justified, T^+ may provide a more efficient test of location for some distributions.

The derived sample data on which these test statistics are based consist of the set of N integer ranks $\{1, 2, \ldots, N\}$ and a corresponding set of N plus and minus signs. The rank i is associated with a plus or minus sign according to the sign of $D_j = X_j - M_0$, where D_j occupies the ith position in the ordered array of absolute differences $|D_j|$. If we let $r(.)$ denote the rank of a random variable, the *Wilcoxon signed-rank statistic* can be written symbolically as

$$T^+ = \sum_{i=1}^{N} Z_i r(|D_i|) \qquad T^- = \sum_{i=1}^{N} (1 - Z_i) r(|D_i|) \tag{7.1}$$

where

$$Z_i = \begin{cases} 1 & \text{if } D_i > 0 \\ 0 & \text{if } D_i < 0 \end{cases}$$

Therefore,

$$T^+ - T^- = 2\sum_{i=1}^{N} r|D_i| - \frac{N(N+1)}{2}$$

Under the null hypothesis, the Z_i are independent and identically distributed Bernoulli random variables with $P(Z_i = 1) = P(Z_i = 0) = 1/2$ so that $E(Z_i) = 1/2$ and $\text{var}(Z_i) = 1/4$. Using the fact that T^+ in (7.1) is a linear combination of these variables, its exact null mean and variance can be determined. We have

$$E(T^+ \mid H_0) = \sum_{i=1}^{N} \frac{r(|D_i|)}{2} = \frac{N(N+1)}{4}$$

Also, since Z_i is independent of $r(|D_i|)$ under H_0 (see Probem 5.25), we can show that

$$\text{var}(T^+ \mid H_0) = \sum_{i=1}^{N} \frac{[r|D_i|]^2}{4} = \frac{N(N+1)(2N+1)}{24} \tag{7.2}$$

A symbolic representation of the test statistic T^+ that is more convenient for the purpose of deriving its mean and variance in general is

$$T^+ = \sum_{1 \leqslant i \leqslant j \leqslant N}\sum T_{ij} \tag{7.3}$$

where

$$T_{ij} = \begin{cases} 1 & \text{if } D_i + D_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

The D_i's are identically distributed under H_0. Now define for all distinct i, j, k the probabilities

$$\begin{aligned} p_1 &= P(D_i > 0) \\ p_2 &= P(D_i + D_j > 0) \\ p_3 &= P(D_i > 0 \text{ and } D_i + D_j > 0) \\ p_4 &= P(D_i + D_j > 0 \text{ and } D_i + D_k > 0) \end{aligned} \tag{7.4}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The moments of the indicator variables for all distinct i, j, k, h are then

$$E(T_{ii}) = p_1 \qquad E(T_{ij}) = p_2$$
$$\text{var}(T_{ii}) = p_1 - p_1^2 \qquad \text{var}(T_{ij}) = p_2 - p_2^2$$
$$\text{cov}(T_{ii}, T_{ik}) = p_3 - p_1 p_2 \qquad \text{cov}(T_{ij}, T_{ik}) = p_4 - p_2^2$$
$$\text{cov}(T_{ij}, T_{hk}) = 0$$

The mean and variance of the linear combination in (7.3) in terms of these moments are

$$E(T^+) = NE(T_{ii}) + \frac{N(N-1)E(T_{ij})}{2} = Np_1 + \frac{N(N-1)p_2}{2} \tag{7.5}$$

$$\begin{aligned}
\text{var}(T^+) &= N\text{var}(T_{ii}) + \binom{N}{2}\text{var}(T_{ij}) + 2N(N-1)\,\text{cov}(T_{ii}, T_{ik}) \\
&\quad + 2N\binom{N-1}{2}\text{cov}(T_{ij}, T_{ik}) + \binom{N}{4}\text{cov}(T_{ij}, T_{hk}) \\
&= Np_1(1-p_1) + \frac{N(N-1)p_2(1-p_2)}{2} \\
&\quad + 2N(N-1)(p_3 - p_1p_2) + N(N-1)(N-2)(p_4 - p_2^2) \\
&= Np_1(1-p_1) + N(N-1)(N-2)(p_4 - p_2^2) \\
&\quad + \frac{N(N-1)}{2}[p_2(1-p_2) + 4(p_3 - p_1p_2)]
\end{aligned} \tag{7.6}$$

The relevant probabilities from (7.4) are now evaluated under the assumption that the population is symmetric and the null hypothesis is true.

$$\begin{aligned}
p_1 &= P(D_i > 0) = 1/2 \\
p_2 &= P(D_i + D_j > 0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_D(u)f_D(v)\,du\,dv \\
&= \int_{-\infty}^{\infty}[1 - F_D(-v)]f_D(v)\,dv \\
&= \int_{-\infty}^{\infty} F_D(v)f_D(v)\,dv = \int_0^1 x\,dx = 1/2 \\
p_3 &= P(D_i > 0 \text{ and } D_i + D_j > 0) \\
&= \int_0^{\infty}\int_{-v}^{\infty} f_D(u)f_D(v)\,du\,dv = \int_0^{\infty}[1 - F_D(-v)]f_D(v)\,dv \\
&= \int_0^{\infty} F_D(v)f_D(v)\,dv = \int_{1/2}^1 x\,dx = 3/8
\end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$
\begin{aligned}
p_4 &= P(D_i + D_j > 0 \text{ and } D_i + D_k > 0) \\
&= P(0 < D_i + D_j < D_i + D_k) + P(0 < D_i + D_k < D_i + D_j) \\
&= 2P(-D_i < D_j < D_k) \\
&= 2\int_{-\infty}^{\infty}\int_{-w}^{\infty}\int_{-v}^{\infty} f_D(u) f_D(v) f_D(w)\, du\, dv\, dw \\
&= 2\int_{-\infty}^{\infty}\int_{-w}^{\infty} [1 - F_D(v)] f_D(v) f_D(w)\, dv\, dw \\
&= 2\int_{-\infty}^{\infty}\int_{-w}^{\infty} f_D(v) f_D(w)\, dv\, dw - 2\int_{-\infty}^{\infty}\int_{-w}^{\infty} F_D(v) f_D(v) f_D(w)\, dv\, dw \\
&= 2\int_{-\infty}^{\infty} [1 - F_D(-w)] f_D(w)\, dw - \int_{-\infty}^{\infty} \{1 - [F_D(-w)]^2\} f_D(w)\, dw \\
&= 2\int_{-\infty}^{\infty} F_D(w)\, dF_D(w) - 1 + \int_{-\infty}^{\infty} [1 - F_D(w)]^2\, dF_D(w) \\
&= 2(1/2) - 1 + (1/3) = 1/3
\end{aligned}
$$

The reader may verify that substitution of these results back in (7.5) and (7.6) gives the mean and variance already found in (7.2).

We use the method described in Chapter 1 to investigate the consistency of T^+.

We can write

$$
E\left[\frac{2T^+}{N(N+1)}\right] = \frac{2p_1}{N+1} + \frac{(N-1)p_2}{N+1}
$$

which equals $\frac{1}{2}$ under H_0 and $\text{var}[2T^+/N(N+1)]$ clearly tends to zero as $N \to \infty$. Therefore, the test with rejection region

$$
T^+ \in R \qquad \text{for } \frac{2T^+}{N(N+1)} - \frac{1}{2} \geqslant k
$$

is consistent against alternatives of the form $p_2 = P(D_1 + D_j > 0) > 0.5$. This result is reasonable since if the true population median exceeds M_0, the sample data would reflect this by having most of the larger ranks correspond to positive differences. A similar two-sided rejection region of T^+ centered on $N(N+1)/4$ is consistent against alternatives with $p_2 \neq 0.5$.

To determine the rejection regions precisely for this consistent test, the probability distribution of T^+ must be determined under the null hypothesis

$$
H_0\colon \theta = P(X > M_0) = 0.5
$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The extreme values of T^+ are zero and $N(N+1)/2$, occurring when all differences are of the same sign, negative or positive, respectively. The mean and variance were found in (7.2). Since T^+ is completely determined by the indicator variables Z_i in (7.1), the sample space can be considered to be the set of all possible N-tuples $\{z_1, z_2, \ldots, z_N\}$ with components either one or zero, of which there are 2^N. Each of these distinguishable arrangements is equally likely under H_0. Therefore, the null probability distribution of T^+ given by

$$P(T^+ = t) = u(t)/2^N \tag{7.7}$$

where $u(t)$ is the number of ways to assign plus and minus signs to the first N integers such that the sum of the positive integers equals t. Every assignment has a conjugate assignment with plus and minus signs interchanged, and T^+ for this conjugate is

$$\sum_{i=1}^{N} i(1 - Z_i) = \frac{N(N+1)}{2} - \sum_{i=1}^{N} iZ_i$$

Since every assignment occurs with equal probability, this implies that the null distribution of T^+ is symmetric about its mean $N(N+1)/4$.

Because of the symmetry property, only one-half of the null distribution need be determined. A systematic method of generating the complete distribution of T^+ for $N = 4$ is shown in Table 7.1.

$$f_{T^+}(t) = \begin{cases} 1/16 & t = 0, 1, 2, 8, 9, 10 \\ 2/16 & t = 3, 4, 5, 6, 7 \\ 0 & \text{otherwise} \end{cases}$$

Tables can be constructed in this way for all N.

To use the signed-rank statistics in hypothesis testing, the entire null distribution is not necessary. In fact, one set of critical values is sufficient for even a two-sided test, because of the relationship

Table 7.1 Enumeration for the distribution of T^+

Value of T^+	*Ranks associated with positive differences*	*Number of sample points u(t)*
10	1,2,3,4	1
9	2,3,4	1
8	1,3,4	1
7	1,2,4; 3,4	2
6	1,2,3; 2,4	2
5	1,4; 2,3	2

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$T^+ + T^- = N(N+1)/2$ and the symmetry of T^+ about $N(N+1)/4$. Large values of T^+ correspond to small values of T^- and T^+ and T^- are identically distributed since

$$\begin{aligned} P(T^+ \geqslant c) &= P\left[T^+ - \frac{N(N+1)}{4} \geqslant c - \frac{N(N+1)}{4}\right] \\ &= P\left[\frac{N(N+1)}{4} - T^+ \geqslant c - \frac{N(N+1)}{4}\right] \\ &= P\left[\frac{N(N+1)}{2} - T^+ \geqslant c\right] \\ &= P(T^- \geqslant c) \end{aligned}$$

Since it is more convenient to work with smaller sums, tables of the left-tailed critical values are generally set up for the random variable T, which may denote either T^+ or T^-. If t_α is the number such that $P(T \leqslant t_\alpha) = \alpha$, the appropriate rejection regions for size α tests of H_0: $M = M_0$ are as follows:

$T^- \leqslant t_\alpha$ for H_1: $M > M_0$

$T^+ \leqslant t_\alpha$ for H_1: $M < M_0$

$T^+ \leqslant t_{\alpha/2}$ or $T^- \leqslant t_{\alpha/2}$ for $H_1 : M \neq M_0$

Suppose that $N = 8$ and critical values are to be found for one- or two-sided tests at nominal $\alpha = 0.05$. Since $2^8 = 256$ and $256(0.05) = 12.80$, we need at least 13 cases of assignments of signs. We enumerate the small values of T^+ in Table 7.2. Since $P(T^+ \leqslant 6) = 14/256 > 0.05$ and $P(T^+ \leqslant 5) = 10/256 = 0.039, t_{0.05} = 5$; the exact probability of a type I error is 0.039. Similarly, we find $t_{0.025} = 3$ with exact $P(T^+ \leqslant 3) = 0.0195$.

Table 7.2 Partial distribution of T_N^+ for $N = 8$

Value of T^+	*Ranks associated with positive differences*	*Number of sample points*
0		1
1	1	1
2	2	1
3	3; 1,2	2
4	4; 1,3	2
5	5; 1,4; 2,3	3
6	6; 1,5; 2,4; 1,2,3	4

When the distribution is needed for several sample sizes, a simple recursive relation can be used to generate the probabilities. Let T_N^+ denote the sum of the ranks associated with positive differences D_i for a sample of N observations. Consider a set of $N-1$ ordered $|D_i|$, with ranks $1, 2, \ldots, N-1$ assigned, for which the null distribution of T_{N-1}^+ is known. To obtain the distribution of T_N^+ from this, an extra observation D_N is added, and we can assume without loss of generality that $|D_N| > |D_i|$ for all $i \leqslant N-1$. The rank of $|D_N|$ is then N. If $|D_N| > 0$, the value of T_N^+ will exceed that of T_{N-1}^+ by the amount N for every arrangement of the $N-1$ observations, but if $|D_N| < 0, T_N^+$ will be equal to T_{N-1}^+. Using the notation in (7.7), this can be stated as

$$P(T_N^+ = k) = \frac{u_N(k)}{2^N} = \frac{u_{N-1}(k-N)P(D_N > 0) + u_{N-1}(k)P(D_N < 0)}{2^{N-1}}$$

$$= \frac{u_{N-1}(k-N) + u_{N-1}(k)}{2^N} \tag{7.8}$$

If N is moderate and systematic enumeration is desired, classification according to the number of positive differences D_i is often helpful. Define the random variable U as the number of positive differences; U follows the binomial distribution with parameter 0.5, so that

$$\begin{aligned} P(T^+ = t) &= \sum_{i=0}^{N} P(U = i \cap T^+ = t) \\ &= \sum_{i=0}^{N} P(U = i) P(T^+ = t \mid U = i) \\ &= \sum_{i=0}^{N} \binom{N}{i} (0.5)^N P(T^+ = t \mid U = i) \end{aligned}$$

A table of critical values and exact significance levels of the Wilcoxon signed-rank test is given in Dunstan, Nix, and Reynolds (1979) for $N \leqslant 50$, and the entire null distribution is given in Wilcoxon, Katti, and Wilcox (1972) for $N \leqslant 50$. Table H of the Appendix of this book gives left-tail and right-tail probabilities of T^+ (or T^-) for $N \leqslant 15$. From a generalization of the central-limit theorem, they asymptotic distribution of T^+ is the normal. Therefore, in the null case, using the moments given in (7.2), the distribution of

$$Z = \frac{4T^+ - N(N+1)}{\sqrt{2N(N+1)(2N+1)/3}} \tag{7.9}$$

approaches the standard normal as $N \to \infty$. The test for, say, H_1: $M > M_0$ can be performed for large N by computing (7.9) and rejecting H_0 for $Z \geqslant z_\alpha$. The approximation is generally adequate for N at least 15. A continuity correction of 0.5 generally improves the approximation.

THE PROBLEM OF ZERO AND TIED DIFFERENCES

Since we assumed originally that the random sample was drawn from a continuous population, the problem of tied observations and zero differences could be ignored theoretically. In practice, generally any zero differences (observations equal to M_O) are ignored and N is reduced accordingly, although the other procedures described for the ordinary sign test in Section 5.4 are equally applicable here. In the case where two or more absolute values of differences are equal, that is, $|d_i| = |d_j|$ for at least one $i \neq j$, the observations are tied. The ties can be dealt with by any of the procedures described in Section 5.6. The midrank method is usually used, and the sign associated with the midrank of $|d_i|$ is determined by the original sign of d_i as before. The probability distribution of T is clearly not the same in the presence of tied ranks, but the effect is generally slight and no correction need be made unless the ties are quite extensive. A thorough comparison of the various methods of treating zeros and ties with this test is given in Pratt and Gibbons (1981).

With large sample sizes when the test is based on the standard normal statistic in (7.9), the variance can be corrected to account for the ties as long as the midrank method is used to resolve the ties. Suppose that t observations are tied for a given rank and that if they were not tied they would be given the ranks $s+1, s+2, \ldots, s+t$. The midrank is then $s + (t+1)/2$ and the sum of squares of these ranks is

$$t\left[s + \frac{(t+1)}{2}\right]^2 = t\left[s^2 + s(t+1) + \frac{(t+1)^2}{4}\right]$$

If these ranks had not been tied, their sum of squares would have been

$$\sum_{i=1}^{t}(s+i)^2 = ts^2 + st(t+1) + \frac{t(t+1)(2t+1)}{6}$$

The presence of these t ties then decreases the sum of squares by

$$\frac{t(t+1)(2t+1)}{6} - \frac{t(t+1)^2}{4} = \frac{t(t+1)(t-1)}{12} \tag{7.10}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Therefore the reduced variance from (7.2) is

$$\operatorname{var}(T^+|H_0) = \frac{N(N+1)(2N+1)}{24} - \frac{\sum t(t^2-1)}{48} \tag{7.11}$$

where the sum is extended over all sets of t ties.

POWER FUNCTION

The distribution of T^+ is approximately normal for large sample sizes regardless of whether the null hypothesis is true. Therefore a large sample approximation to the power can be calculated using the mean and variance given in (7.5) and (7.6). The distribution of $X - M_0$ under the alternative would need to be specified in order to calculate the probabilities in (7.4) to substitute in (7.5) and (7.6).

The asymptotic relative efficiency of the Wilcoxon signed-rank test relative to the t test is at least 0.864 for any distribution continuous and symmetric about zero, is 0.955 for the normal distribution, and is 1.5 for the double exponential distribution.

It should be noted that the probability distribution of T^+ is not symmetric when the null hypothesis is not true. Further, T^+ and T^- are not identically distributed when the null hypothesis is not true. We can still find the probability distribution of T^- from that of T^+, however, using the relationship

$$P(T^- = k) = P\left[\frac{N(N+1)}{2} - T^+ = k\right] \tag{7.12}$$

SIMULATED POWER

Calculating the power of the signed-rank test, even using the normal approximation, requires a considerable amount of work. It is much easier to simulate the power of the test, as we did for the sign test in Section 5.4. Again we use a MINITAB Macro program for the calculations and in order to compare the results with those obtained for the sign test, we use $N = 13, \alpha = 0.05, M_0 = 0.5$ and $M_1 = 0.6256$.

Simulating the power of the signed-rank test consists of the following steps. First, we determine the rejection region of the signed-rank test from Table H of the Appendix as $T^+ \geqslant 70$ with exact $\alpha = 0.047$. We generate 1000 random samples each of size $N = 13$ from a normal distribution with mean 0.6256 and variance 1 and calculate the signed-rank statistic T^+ for each. For each of these statistics we check to see if it exceeds the critical value 70 or not. Finally, we count the number of times, out of 1000, that the signed-rank test rejects the null hypothesis and divide this number by 1000. This gives a

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

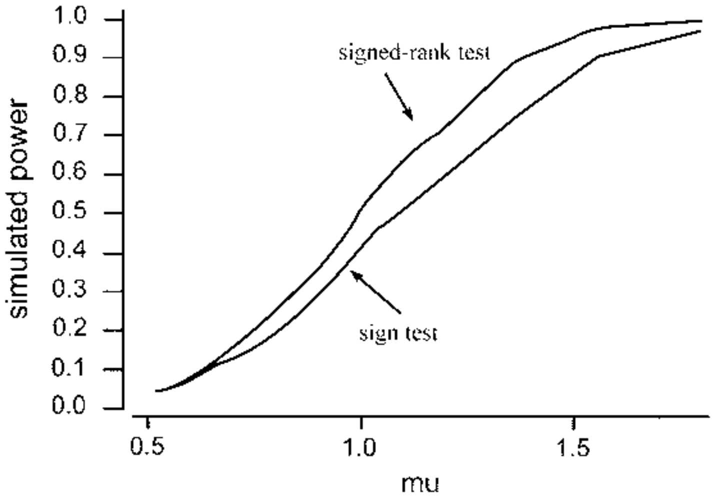


Fig. 7.1 Simulated power of the sign and the signed-rank rank test for the normal distribution.

simulated (estimated) value of the power of the sign test with $N = 13, \alpha = 0.0461, M_0 = 0.50, M_1 = 0.6256$. The program code is shown below. Note that the program also calculates the simulated power of the sign test and plots the two simulated power curves on the same graph. This graph is shown in Figure 7.1.

```
macro signrank
# simulates powers of sign and the signed-rank test
signrank
mcolumn c1 c2 c3 c4 pow1 pow2 theta phiinv mu
mconstant k1 k2 k3 k4 k5 cow mu1
let k5=1
set theta
(.5:.9/.05)1
end
#print theta
invcdf theta phiinv;
normal 0.0 1.0.
let mu=.5+phiinv
mlable 4
let cow=1
let pow1 (k5)=0
let pow2 (k5)=0
```

```
mlabel 1
let mu1=mu(k5)
random 13 c1;
normal mu1 1.
let c1=c1-.-50
# calculate Sign and Signed Rank
let c2=(c1 gt 0)
let k1=sum(c2)
# k1=value of Sign stat
let c3=abs(c1)
let c3=rank(c3)
let c4=c2*c3
let k2=sum(c4)
# k2=value of Signed-Rank stat
# print c1 c2 c3 c4 k1 k2
# k4 is the critical value of sign test from Table G
# at n=13,exact alpha=0.0461
let k3=10
# k4 is the critical value for the signed
# rank test from Table H at n=13, exact alpha=0.47
let k4=70
if k1 ge 10
let pow1(k5)=pow1(k5)+1
else
let pow1(k5)=pow1(k5)+0
endif
if k2 ge 70
let pow2(k5)=pow2(k5)+1
else
let pow2(k5)=pow2(k5)+0
endif
let cow=cow+1
if cow gt 1000
go to 2
else
go to 1
endif
mlabel 2
let pow1(k5)=pow1(k5)/1000
let pow2(k5)=pow2(k5)/1000
print k5 mu(k5)theta(k5) pow1(k5)pow2(k5)
let k5=k5+1
if k5 gt 9
go to 3
else
go to 4
endif
```

```
mlabel 3
plot1*mu pow2*mu;
 connect;
 color 1 2;
axis 2;
 label 'simulated power';
overlay.
endmacro
```

The output from the MACRO is shown below; pow1 and pow2 are, respectively, the computed powers of the sign and the signed-rank test, based on 1000 simulations.

mu	theta	pow1	pow2
0.50000	0.50	0.033	0.036
0.62566	0.55	0.095	0.100
0.75335	0.60	0.166	0.211
0.88532	0.65	0.292	0.352
1.02440	0.70	0.465	0.568
1.17449	0.75	0.582	0.710
1.34162	0.80	0.742	0.885
1.53643	0.85	0.897	0.974
1.78155	0.90	0.968	0.994

SAMPLE SIZE DETERMINATION

In order to make an inference regarding the population median using the signed-rank test, we need to have a random sample of observations. If we are allowed to choose the sample size, we might want to determine the value of N such that the test has size α and power $1-\beta$, given the null and the alternative hypotheses and other necessary assumptions. Recall that for the sign test against the one- sided upper-tailed alternative, we solved for N such that

$$\text{Size} = \sum_{i=k_\alpha}^{N} \binom{N}{i} (0.5)^N \leqslant \alpha$$

$$\text{and} \qquad \text{power} = \sum_{i=k_\alpha}^{N} \binom{N}{i} \theta^i (1-\theta)^{N-i} \geqslant 1-\beta$$

where α, $1-\beta$, and $\theta = P(X > M_0|H_1)$ are all specified. We noted there that the solution is much easier to obtain using the normal approx-

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

imation; the same is true for the Wilcoxon signed-rank test, as we now illustrate. Note that the theory is presented here in terms of the signed-rank statistic T^+ but the same approach will hold for any test statistic whose distribution can be approximated by a normal distribution under both the null and the alternative hypotheses.

Under the normal approximation, the power of a size α signed-rank test against the alternative $H_1 : M > M_0$ is $P(T^+ \geqslant \mu_0 + z_\alpha \sigma_0 | H_1)$, where μ_0 and σ_0 are, respectively, the null mean and the null standard deviation of T^+. It can be easily shown (see Noether, 1987) that this power equals a specified $1-\beta$ if

$$\left(\frac{\mu_0 - \mu}{\sigma}\right)^2 = (z_\alpha + \rho z_\beta)^2 \tag{7.13}$$

where μ and σ are, respectively, the mean and the standard deviation of T^+ under the alternative hypothesis. We denote the relation between standard deviations by $\rho = \sigma/\sigma_0$. Since σ is unknown and is difficult to evaluate [see (7.6)], ρ is unknown. One possibility is to take ρ equal to 1 and this is what is done; such an assumption is reasonable for alternative hypotheses that are not too different from the null hypothesis.

If we substitute the expressions for μ_0, σ_0, and μ [see (7.5)] into (7.13), we need to solve for N in

$$\frac{[N(p_1 - 0.5) + (N(N-1)(p_2 - 0.5))/2]^2}{N(N+1)(2N+1)/24} = (z_\alpha + z_\beta)^2 \tag{7.14}$$

Note that $p_1 = P(X_i > M_0)$ and $p_2 = P(X_i + X_j > 2M_0)$ under the alternative $H_1 : M > M_0$. The sample size calculations from (7.14) are shown in Table 7.3 for $\alpha = 0.05$, $1-\beta = 0.95$, assuming the underlying distribution is standard normal. These calculations are the solution for N in (7.14) done in EXCEL using the solver application. Note that the answer for the sample size N, shown in the fifth column, needs to be rounded up to that next larger integer. Thus, for example, assuming normality and the $M_0 = 0$, $M_1 = 0.5$, $\alpha = 0.05$, we need to have approximately 33 observations in our sample for a power of 0.95 and a one-sided alternative.

A similar derivation can be used to find a sample size formula when the alternative is two-sided. The details are left to the reader as an exercise.

It may be noted that the sample size formula in (7.14) is not distribution-free since it depends on the underlying distribution

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 7.3 Calculations for sample size determination in EXCEL

M_0	M_1	P_1	P_2	N	$p_1 - 0.5$	$0.5N(N-1)(p_2-0.5)$	$N(N+1)(2N+1)/24$	*Error*
0	0.1	0.539828	0.57926	575.7201	0.039828	13112.63915	15943497	9.83E-11
0	0.2	0.57926	0.655422	150.7868	0.07926	1755.166447	288547.2	−9.6E-09
0	0.3	0.617911	0.725747	72.17494	0.117911	579.8362509	31985.43	−2.1E-07
0	0.4	0.655422	0.788145	44.75747	0.155422	282.1618931	7723.9	−2.7E-08
0	0.5	0.691462	0.841345	32.17465	0.191462	171.190114	2906.364	−1.7E-07
0	0.6	0.725747	0.88493	25.45262	0.225747	119.7870853	1456.134	−4.6E-08
0	0.7	0.758036	0.919243	21.51276	0.258036	92.50310853	888.4195	−5E-08
0	0.8	0.788145	0.945201	19.06405	0.288145	76.65779392	623.6078	−2.1E-08
0	0.9	0.81594	0.96407	17.48421	0.31594	66.87552029	484.3471	−6.1E-09
0	1.0	0.841345	0.97725	16.44019	0.341345	60.57245533	404.7574	−2.6E-09

through the parameters p_1 and p_2. Noether (1987) proposed approximating the left-hand side of (7.14) as $3N(p_2 - 0.5)^2$ and solving for N, which yields

$$N = \frac{(z_\alpha + z_\beta)^2}{3(p_2 - 0.5)^2} \tag{7.15}$$

This formula still depends on p_2; Noether (1987) suggested a choice for this parameter in terms of an "odds-ratio." The reader is referred to his paper for details.

For a two-sided test, we can use (7.15) with α replaced by $\alpha/2$.

We illustrate the use of (7.15) for this example where $\alpha = 0.05$, $1-\beta = 0.95$. If $M_1 = 0.1$ and $p_2 = 0.579$, we find $N = 578.12$ from (7.15); if $M_1 = 1.0$ and $p_2 = 0.977$, we find $N = 15.86$ from (7.15). The corresponding values shown in Table 7.1 are $N = 575.72$ and $N = 16.44$, respectively.

CONFIDENCE-INTERVAL PROCEDURES

As with the ordinary one-sample sign test, the Wilcoxon signed-rank procedure lends itself to confidence-interval estimation of the unknown population median M. In fact, two methods of interval estimation are available here. Both will give the confidence limits as those values of M which do not lead to rejection of the null hypothesis, but one amounts to a trial-and-error procedure while the other is systematic and provides a unique interval. For any sample size N, we can find that number $t_{\alpha/2}$ such that if the true population median is M and T is calculated for the derived sample values $X_i - M$, then

$$P(T^+ \leqslant t_{\alpha/2}) = \frac{\alpha}{2} \qquad \text{and} \qquad P(T^- \leqslant t_{\alpha/2}) = \frac{\alpha}{2}$$

The null hypothesis will not be rejected for all numbers M which make $T^+ > t_{\alpha/2}$ and $T^- > t_{\alpha/2}$. The confidence interval technique is to find those two numbers, say M_1 and M_2 where $M_1 < M_2$, such that when T is calculated for the two sets of differences $X_i - M_1$ and $X_i - M_2$, at the significance level α, T^+ or T^-, whichever is smaller, is just short of significance, i.e., slightly larger than $t_{\alpha/2}$. Then the $100(1-\alpha)$ percent confidence-interval estimate of M is $M_1 < M < M_2$.

In the trial-and-error procedure, we simply choose some suitable values of M and calculate the resulting values of T^+ or T^-, stopping whenever we get numbers slightly larger than $t_{\alpha/2}$. This generally does not lead to a unique interval, and the manipulations can be tedious even for moderate sample sizes. The technique is best illustrated by an

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

example. The following eight observations are drawn from a continuous, symmetric population:

$$-1, 6, 13, 4, 2, 3, 5, 9 \tag{7.16}$$

For $N = 8$ the two-sided rejection region of nominal size 0.05 was found earlier by Table 7.2 to be $t_{\alpha/2} = 3$ with exact significance level

$$\alpha = P(T^+ \leqslant 3) + P(T^- \leqslant 3) = 10/256 = 0.039$$

In Table 7.4 we try six different values for M and calculate T^+ or T^-, whichever is smaller, for the differences $X_i - M$. The example illustrates a number of difficulties which arise. In the first trial choice of M, the number 4 was subtracted and the resulting differences contained three sets of tied pairs and one zero even though the original sample contained neither ties nor zeros. If the zero difference is ignored, N must be reduced to 7 and then the $t_{\alpha/2} = 3$ is no longer accurate for $\alpha = 0.039$. The midrank method could be used to handle the ties, but this also disturbs the accuracy of $t_{\alpha/2}$. Since there seems to be no real solution to these problems, we try to avoid zeros and ties by judicious choices for our M values for subtraction. Since these data are all integers, a choice for M which is noninteger valued obviously reduces the likelihood of ties and makes zero values impossible. Since T^- for the differences $X_i - 1.1$ yields $T^- = 3.5$ using the midrank method, we will choose $M_1 = 1.5$. The next three columns represent an attempt to find an M which makes T^+ around 4. These calculations illustrate the fact that M_1 and M_2 are far from being unique. Clearly M_2 is in the vicinity of 9, but the differences $X_i - 9$ yield a zero. We conclude there is no need to go further. An approximate 96.1 percent confidence

Table 7.4 Trial-and-error determination of endpoints

X_I	$X_i - 4$	$X_i - 1.1$	$X_i - 1.5$	$X_i - 9.1$	$X_i - 8.9$	$X_i - 8.95$
−1	−5	−2.1	−2.5	−10.1	−9.9	−9.95
6	2	4.9	4.5	−3.1	−2.9	−2.95
13	9	11.9	11.5	3.9	4.1	4.05
4	0	2.9	2.5	−5.1	−4.9	−4.95
2	−2	0.9	0.5	−7.1	−6.9	−6.95
3	−1	1.9	1.5	−6.1	−5.9	−5.95
5	1	3.9	3.5	−4.1	−3.9	−3.95
9	5	7.9	7.5	−0.1	0.1	0.05
T^+ or T^-		3	3.5	3	5	5

interval on M is given by $1.5 < M < 9$. The interpretation is that hypothesized values of M within this range will lead to acceptance of the null hypothesis for an exact significance level of 0.039.

This procedure is undoubtedly tedious, but the limits obtained are reasonably accurate. The numbers should be tried systematically to narrow down the range of possibilities. Thoughtful study of the intermediate results usually reduces the additional number of trials required.

A different method of construction which leads to a unique interval and is much easier to apply is described in Noether [(1967), pp. 57–58]. The procedure is to convert the interval $T^+ > t_{\alpha/2}$ and $T^- > t_{\alpha/2}$ to an equivalent statement on M whose end points are functions of the observations X_i. For this purpose we must analyze the comparisons involved in determining the ranks of the differences $r(|X_i - M_0|)$ and the signs of the differences $X_i = M_0$ since T^+ and T^- are functions of these comparisons. Recall from (5.1) that the rank of any random variable in a set $\{V_1, V_2, \ldots, V_N\}$ can be written symbolically as

$$r(V_i) = \sum_{k=1}^{N} S(V_i - V_k) = \sum_{k \neq i} S(V_i - V_k) + 1$$

where

$$S(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{if } u \leqslant 0 \end{cases}$$

To compute a rank, then we make $\binom{N}{2}$ comparisons of pairs of different numbers and one comparison of a number with itself. To compute the sets of all ranks, we make $\binom{N}{2}$ comparisons of pairs and N identity comparisons, a total of $\binom{N}{2} + N = N(N+1)/2$ comparisons. Substituting the rank function in (7.1), we obtain

$$\begin{aligned} T^+ &= \sum_{i=1}^{N} Z_i \, r(|X_i - M_0|) \\ &= \sum_{i=1}^{N} Z_i + \sum_{i=1}^{N} \sum_{k \neq i} Z_i \, S(|X_i - M_0| - |X_k - M_0|) \end{aligned} \tag{7.17}$$

Therefore these comparisons affect T^+ as follows:

1. A comparison of $|X_i - M_0|$ with itself adds 1 to T^+ if $X_i - M_0 > 0$.
2. A comparison of $|X_i - M_0|$ with $|X_k - M_0|$ for any $i \neq k$ adds 1 to T^+ if $|X_i - M_0| > |X_k - M_0|$ *and* $X_i - M_0 > 0$, that is, $X_i - M_0 > |X_k - M_0|$. If $X_k - M_0 > 0$, this occurs when $X_i > X_k$, and if $X_k - M_0 < 0$, we have $X_i + X_k > 2M_0$ or $(X_i + X_k)/2 > M_0$. But when $X_i - M_0 > 0$ and $X_k - M_0 > 0$, we have $(X_i + X_k)/2 > M_0$ also.

Combining these two results, then, $(X_i + X_k)/2 > M_0$ is a necessary condition for adding 1 to T^+ for all i, k. Similarly, if $(X_i + X_k)/2 < M_0$, then this comparison adds 1 to T^-. The relative magnitudes of the $N(N+1)/2$ averages of pairs $(X_i + X_k)/2$ for all $i \leqslant k$, called the *Walsh averages*, then determine the range of values for hypothesized numbers M_0 which will not lead to rejection of H_0. If these $N(N+1)/2$ averages are arranged as order statistics, the two numbers which are in the $(t_{\alpha/2} + 1)$ position from either end are the endpoints of the $100(1-\alpha)$ percent confidence interval on M. Note that this procedure is exactly analogous to the ordinary sign-test confidence interval except that here the order statistics are for the averages of all pairs of observations instead of the original observations.

The data in (7.16) for $N = 8$ arranged in order of magnitude are $-1, 2, 3, 4, 5, 6, 9, 13$, and the 36 Walsh averages are given in Table 7.5. For exact $\alpha = 0.039$, we found before that $t_{\alpha/2} = 3$. Since the fourth largest numbers from either end are 1.5 and 9.0, the confidence interval is $1.5 < M < 9$ with exact confidence coefficient $\gamma = 1 - 2(0.039) = 0.922$. This result agrees exactly with that obtained by the previous method, but this will not always be the case since the trial-and-error procedure does not yield unique endpoints.

The process of determining a confidence interval on M by the above method is much facilitated by using the graphical method of construction, which can be described as follows. Each of the N ob-

Table 7.5 Walsh averages for data in (7.16)

−1.0	0.5	1.0	1.5	2.0	2.5	4.0	6.0
2.0	2.5	3.0	3.5	4.0	5.5	7.5	
3.0	3.5	4.0	4.5	6.0	8.0		
4.0	4.5	5.0	6.5	8.5			
5.0	5.5	7.0	9.0				
6.0	7.5	9.5					
9.0	11.0						
13.0							

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

servations x_i is denoted by a dot on a horizontal scale. The closed interval $[X_{(1)}, X_{(N)}]$ then includes all dots. Form an isosceles triangle ABC by lines joining $x_{(1)}$ at A and $x_{(N)}$ at B each with a point C anywhere on the vertical line passing through the midrange value $(x_{(1)} + x_{(N)})/2$. Through each point x_i on the line segment AB draw lines parallel to AC and BC, marking each intersection with a dot. There will be $N(N+1)/2$ intersections, the abscissas of which are all the $(x_i + x_k)/2$ values where $1 \leqslant i \leqslant k \leqslant N$. Vertical lines drawn through the $(t_{\alpha/2} + 1)$st intersection point from the left and right will allow us to read the respective confidence-interval end points on the horizontal scale. Figure 7.2 illustrates this method for the numerical data above.

PAIRED-SAMPLE PROCEDURES

The Wilcoxon signed-rank test was actually proposed for use with paired-sample data in making inferences concerning the value of the median of the population of differences. Given a random sample of N pairs

$$(X_1,Y_1),\ (X_2,Y_2), \ldots, (X_N,Y_N)$$

their differences are

$$X_1 - Y_1, X_2 - Y_2, \ldots, X_N - Y_N$$

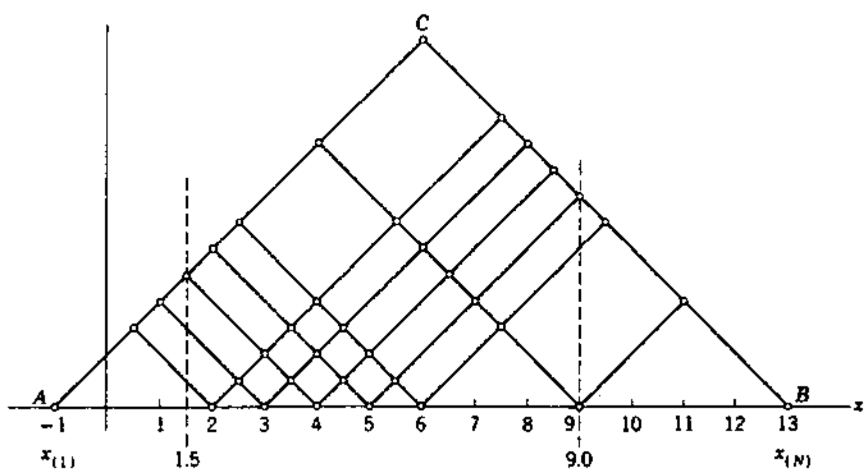


Fig. 7.2 Graphical determination of confidence interval.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

We assume these are independent observations from a population of differences which is continuous and symmetric with median M_0. In order to test the hypothesis

$$H_0\colon M_D = M_0$$

form the N differences $D_i = X_i - Y_i - M_0$ and rank their absolute magnitudes from smallest to largest using integers $\{1, 2, \ldots, N\}$, keeping track of the original sign of each difference. Then the above procedures for hypothesis testing and confidence intervals are equally applicable here with the same notation, except that the parameter M_{D} must be interpreted now as the median of the population of differences.

USE OF WILCOXON STATISTICS TO TEST FOR SYMMETRY

The Wilcoxon signed-rank statistics can also be considered tests for symmetry if the only assumption made is that the random sample is drawn from a continuous distribution. If the null hypothesis states that the population is symmetric with median M_0, the null distributions of T^+ and T^- are exactly the same as before. If the null hypothesis is accepted, we can conclude that the population is symmetric and has median M_0. On the other hand, if the null hypothesis is rejected, we cannot tell which portion (or all) of the composite statement is not consistent with the sample outcome. With a two-sided alternative, for example, we must conclude that either the population is symmetric with median not equal to M_0, or the population is asymmetric with median equal to M_0, or the population is asymmetric with median not equal to M_0. Such a broad conclusion is generally not satisfactory, and this is why in most cases the assumptions that justify a test procedure are separated from the statement of the null hypothesis.

APPLICATIONS

The appropriate rejection regions and P values for T^+, called the sum of the positive ranks, are given below. Note that t is the observed value of T^+.

Alternative	*Exact rejection region*	*Exact P-value*
$M > M_0$	$T^+ \geqslant t_\alpha$	$P(T^+ \geqslant t \mid H_0)$
$M < M_0$	$T^+ \leqslant t'_\alpha$	$P(T^+ \leqslant t \mid H_0)$
$M \neq M_0$	$T^+ \leqslant t'_{\alpha/2}$ or $T^+ \geqslant t_{\alpha/2}$	2(smaller of the above)

Table H gives the distribution of T^+ for $N \leqslant 15$ as left-tail probabilities for $T^+ \leqslant N(N+1)/4$ and right-tail for $T^+ \geqslant N(N+1)/4$. This table can be used to find exact critical values for a given α or to find exact P values. For $N > 15$, the appropriate rejection regions and the P values based on the normal approximation with a continuity correction are as follows:

Alternative	*Approximate rejection region*	*Approximate P value*
$M > M_0$	$T^+ \geqslant \frac{N(N+1)}{4} + 0.5 + z_\alpha\sqrt{\frac{(N+1)(2N+1)}{24}}$	$1 - \Phi\left[\frac{t - 0.5 - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}\right]$
$M < M_0$	$T^+ \leqslant \frac{N(N+1)}{4} - 0.5 - z_\alpha\sqrt{\frac{(N+1)(2N+1)}{24}}$	$\Phi\left[\frac{t + 0.5 - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}\right]$
$M \neq 0$	Both above with $z_{\alpha/2}$	2(smaller of the above)

If ties are present, the variance term in these rejection regions should be replaced by (7.11).

The corresponding confidence interval estimate of the median has endpoints which are $(t_{\alpha/2}+1)^{\text{st}}$ from the smallest and largest of the Walsh averages, where $t_{\alpha/2}$ is the left-tail critical value in Table H for the given N. The choice of exact confidence levels is limited to $1-2P$ where P is a tail probability in Table H. Therefore the critical value $t_{\alpha/2}$ is the left-tail table entry corresponding to the chosen P. Since the entries are all of the nonnegative integers, $(t_{\alpha/2}+1)$ is the rank of $t_{\alpha/2}$ among the table entries for that N.

Thus, in practice, the confidence interval endpoints are the uth smallest and uth largest of the $N(N+1)/2$ Walsh averages $W_{ik} = (X_i + X_k)/2$ for all $1 \leqslant i,\, k \leqslant N$, or

$$W_{(u)} \leqslant M \leqslant W_{[N(N+1)/2-u+1]}$$

The appropriate value of u for confidence $1-2P$ is the rank of that left-tail P among the entries in Table H for the given N. For $N > 15$, we find u from

$$u = \frac{N(N+1)}{4} + 0.5 - z_{\alpha/2}\sqrt{\frac{N(N+1)(2N+1)}{24}}$$

and round down to the next smaller integer if the result is not an integer. If zeros or ties occur in the averages, they should all be counted in determining the endpoints.

These Wilcoxon signed-rank test procedures are applicable to paired samples in exactly the same manner as long as X is replaced by

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

the differences $D = X - Y$ and M is interpreted as the median M_D of the distribution of $X - Y$.

As in the case of the sign test, the confidence-interval estimate of the median or median difference can be based on all N observations even if there are zeros and/or ties. Thus a hypothesis test concerning a value for the median or median difference when the data contain zeros and/or ties will be more powerful if the decision is based on the confidence-interval estimate rather than on a hypothesis test procedure.

Example 7.1 A large company was disturbed about the number of person-hours lost per month due to plant accidents and instituted an extensive industrial safety program. The data below show the number of person-hours lost in a month at each of eight different plants before and after the safety program was established. Has the safety program been effective in reducing time lost from accidents? Assume the distribution of differences is symmetric.

Plant	*Before*	*After*
1	51.2	45.8
2	46.5	41.3
3	24.1	15.8
4	10.2	11.1
5	65.3	58.5
6	92.1	70.3
7	30.3	31.6
8	49.2	35.4

Solution Because of the symmetry assumption, we can use the Wilcoxon signed-rank test instead of the sign test on these data. We take the differences $D =$ Before minus After and test $H_0\colon M_D = 0$ versus $H_1\colon M_D > 0$ since the program is effective if these differences are large positive numbers. Then we rank the absolute values and sum the positive ranks. The table below shows these calculations.

Plant	D	$\lvert D \rvert$	$r(\lvert D \rvert)$
1	5.4	5.4	4
2	5.2	5.2	3
3	8.3	8.3	6
4	−0.9	0.9	1
5	6.8	6.8	5
6	21.8	21.8	8
7	−1.3	1.3	2
8	13.8	13.8	7

We have $T^+ = 33$ and Table H for $N = 8$ gives the right-tail probability as 0.020. The program has been effective at the 0.05 level.

The following computer printouts illustrate the solution to Example 7.1 using the MINITAB, STATXACT and SAS packages.

```
***********************************
   MINITAB SOLUTION TO EXAMPLE 7.1
***********************************

Wilcoxon Signed Rank Test: B-A

Test of median = 0.000000 versus median  >  0.000000

          N for    Wilcoxon              Estimated
      N    Test   Statistic        P        Median
B-A   8       8        33.0    0.021         6.600

***********************************
   STATXACT SOLUTION TO EXAMPLE 7.1
***********************************

Wilcoxon Signed Rank Test

Summary of Exact distribution of WILCOXON SIGNED RANK statistic:
   Min       Max      Mean     Std-dev   Observed    Standardized
0.0000     36.00     18.00       7.141      33.00           2.100

Asymptotic Inference:
   One-sided p-value: Pr {   Test Statistic .GE. Observed }  =  0.0178
   Two-sided p-value: 2 * One-sided                          =  0.0357

Exact Inference:
  One-sided p-value: Pr {   Test Statistic .GE. Observed }   =  0.0195
                     Pr {   Test Statistic .EQ. Observed }   =  0.0078
  Two-sided p-value: Pr { | Test Statistic - Mean |
                            .GE.   | Observed - Mean |       =  0.0391
  Two-sided p-value: 2*One-Sided                             =  0.039

***************************
SAS SOLUTION TO EXAMPLE 7.1
***************************
Program:

DATA EX631;
   INPUT BEFORE AFTER;
  DIFF=BEFORE-AFTER;
```

```
DATALINES;
51.2   45.8
46.5   41.3
24.1   15.8
10.2   11.1
65.3   58.5
92.1   70.3
30.3   31.6
49.2   35.4
;
PROC UNIVARIATE DATA=EX631;
       VAR DIFF;
RUN;
```

Output:

```
           Test              -Statistic-      -----p Value------

     Student's t         t  2.754154     Pr >  |t|     0.0283
     Sign                M         2     Pr >= |M|     0.2891
     Signed Rank         S        15     Pr >= |S|     0.0391
```

The MINITAB solution uses the normal approximation with a continuity correction. The STATXACT solution gives the asymptotic results based on the normal approximation without a continuity correction. Only a portion of the output from SAS PROC UNIVARIATE is shown. This output provides a lot of information, including important descriptive statistics such as the sample mean, variance, interquartile range, etc., which are not shown. Note that the SAS signed-rank statistic is calculated as $T^+ - n(n+1)/4 = 33 - 18 = 15$ (labeled S) and the P value given is two-tailed. The required one-tailed P value can be found as 0.0391/2 = 0.1955, which agrees with other calculations. It is interesting that for these data both the t-test and the signed-rank test clearly lead to a rejection of the null hypothesis at the 0.05 level of significance but the sign test does not.

Example 7.2 Assume the data in Example 7.1 come from a symmetric distribution and find a 90% confidence-interval estimate of the median difference, computed as After minus Before.

Solution Table H for $N = 6$ shows that $P = 0.047$ for confidence $1 - 2(0.047) = 0.906$, and 0.047 has rank three in Table H so that $u = 3$. Thus the 90.6% confidence-interval endpoints for the median difference are the third smallest and third largest Walsh averages.

The $6(7)/2 = 21$ Walsh averages of differences $(D_i + D_k)/2$ are shown in the table below.

−2.0	−1.0	1.0	3.0	4.0	8.0
−1.5	0.0	2.0	3.5	6.0	
−0.5	1.0	2.5	5.5		
0.5	1.5	4.5			
1.0	3.5				
3.0					

So the third smallest and third largest Walsh averages are −1.0 and 5.5, respectively and the 90.6% confidence-interval for the median difference is (−1.0, 5.5). Note that by listing the After minus Before data in an array across the top row of this table of Walsh averages, identification of the confidence-interval endpoints is greatly simplified.

The MINITAB and STATXACT solutions to this example are shown below. The MINITAB solution agrees exactly with our hand calculations. The STATXACT solution gives an asymptotic interval that agrees with our exact solution; the interval labeled exact uses the second smallest and the second largest Walsh averages, which provides the 93.8% confidence interval.

```
                ***********************************
                  MINITAB SOLUTION TO EXAMPLE 7.2
                ***********************************

Wilcoxon Signed Rank CI: C1

                   Estimated    Achieved
              N       Median  Confidence  Confidence Interval
C1            6         2.00        90.7  (    -1.00,     5.50)

                 *********************************
                 STATXACT SOLUTION TO EXAMPLE 7.2
                 *********************************

HODGES-LEHMANN ESTIMATES OF MEDIAN DIFFERENCE
Summary of Exact distribution of WILCOXON SIGNED RANK

    Min     Max     Mean     Std-dev     Observed  Standardized
 0.0000   21.00    10.50       4.757        16.50         1.261

Point Estimate of Median Difference : Lambda  =         2.000

 90.00% Confidence Interval for Lambda :
           Asymptotic : (      -1.000 ,       5.500)
           Exact      : (      -1.500 ,       6.000)
```


MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

5.8 SUMMARY

In this chapter we presented the procedures for hypothesis tests and confidence interval estimates for the pth quantile of any continuous distribution for any specified p, $0 < p < 1$, based on data from one sample or paired samples. These procedures are all based on using the pth sample quantile as a point estimate of the pth population quantile and use the binomial distribution; they have no parametric counterparts. The sample quantiles are all order statistics of the sample. Other estimates of the population quantiles have been introduced in the literature; most of these are based on linear functions of order statistics, say $\sum a_i X_{(i)}$. The one proposed by Harrell and Davis (1982) has been shown to be better than ours for a wide variety of distributions. Dielman, Lowry and Pfaffenberger (1994) present a Monte Carlo comparison of the performance of various sample quantile estimators for small sample sizes.

The pth quantile when $p = 0.5$ is the median of the distribution and we have inference procedures based on the sign test in Section 5.4 and the Wilcoxon signed-rank test in Section 5.7. Both tests are generally useful in the same experimental situations regarding a single sample or paired samples. The assumptions required are minimal – independence of observations and a population which is continuous at M for the ordinary sign test and continuous everywhere and symmetric for the Wilcoxon signed-rank test. Experimentally, both tests have the problem of zero differences, and the Wilcoxon test has the additional problem of ties. Both tests are applicable when quantitative measurements are impossible or not feasible, as when rating scales or preferences are used. For the Wilcoxon test, information concerning relative magnitudes as well as directions of differences is required. Only the sign test can be used for strictly dichotomous data, like yes-no observations. Both are very flexible and simple to use for hypothesis testing or constructing confidence intervals. The null distribution of the sign test is easier to work with since binomial tables are readily available. The normal approximation is quite accurate for even moderate N in both cases, and neither is particularly hampered by the presence of a moderate number of zeros or ties.

For hypothesis testing, in the paired-sample case the hypothesis need not state an actual median difference but only a relation between medians if both populations are assumed symmetric. For example, we might test the hypothesis that the X population values are on the average p percent larger than Y values. Assuming the medians are a reliable indication of size, we would write

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$H_0: M_X = (1 + 0.01p)M_Y$$

and take differences $D_i = X_i - (1 + 0.01p)Y_i$ and perform either test on these derived data as before.

Both tests have a corresponding procedure for finding a confidence interval estimate of the median of the population in the one-sample case and the median difference in the paired-sample case. We have given expressions for sample size determination and power calculations.

Only the Wilcoxon signed-rank statistics are appropriate for tests of symmetry since the ordinary sign-test statistic is not at all related to the symmetry or asymmetry of the population. We have $P[(X_i - M) > 0] = 0.5$ always, and the sole criterion of determining K in the sign test is the number of positive signs, thus ignoring the magnitudes of the plus and minus differences. There are other extensions and modifications of the sign-test type of criteria [see, for example, Walsh (1949*a*,*b*)].

If the population is symmetric, both sign tests can be considered to be tests for location of the population mean and are therefore direct nonparametric counterparts to Student's t test. As a result, comparisons of their performance are of interest. As explained in Chapter 1, one way to compare performance of tests is by computing their asymptotic relative efficiency (ARE) under various distribution assumptions. The asymptotic relative efficiency of the ordinary sign test relative to the t test is $2/\pi = 0.637$, and the ARE of the Wilcoxon signed-rank test relative to the t test is $3/\pi = 0.955$, both calculated under the assumption of normal distributions. How these particular results were obtained will be discussed in Chapter 13. It is not surprising that both ARE values are less than one because the t test is the best test for normal distributions. It can be shown that the ARE of the Wilcoxon signed-rank test is always at least 0.864 for any continuous symmetric distribution, whereas the corresponding lower bound for the ordinary sign test is only 1/3. The ARE of the sign test relative to the Wilcoxon signed-rank test is 2/3 for the normal distribution and 1/3 for the uniform distribution. However, the result is 4/3 for the double exponential distribution; the fact that this ARE is greater than one means that the sign test performs better than the signed-rank test for this particular symmetric but heavy-tailed distribution. Similarly, the Wilcoxon signed-rank test performs better than the t test for some nonnormal distributions; for example, the ARE is 1.50 for the double exponential distribution and 1.09 for the logistic distribution, which are both heavy-tailed distributions.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

PROBLEMS

5.1. Give a functional definition similar to (5.1) for the rank $r(X_i)$ of a random variable in any set of N independent observations where ties are dealt with by the midrank method. *Hint*: In place of $S(u)$ in (5.2), consider the function

$$c(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1/2 & \text{if } u = 0 \\ 1 & \text{if } u > 0 \end{cases}$$

5.2. Find the correlation coefficient between variate values and ranks in a random sample of size N from

(a) The uniform distribution

(b) The standard normal distribution

(c) The exponential distribution

5.3. Verify the cumulative distribution function of differences given in (4.14) and the result $M = -2 + \sqrt{3}$. Find and graph the corresponding probability function of differences.

5.4. Answer parts (a) through (e) using (i) the sign-test procedure and (ii) the Wilcoxon signed-rank test procedure.

(a) Test at a significance level not exceeding 0.10 the null hypothesis H_0: $M = 2$ against the alternative H_1: $M > 2$, where M is the median of the continuous symmetric population from which is drawn the random sample:

$$-3, -6, 1, 9, 4, 10, 12$$

(b) Give the exact probability of a type I error in (a)

(c) On the basis of the following random sample of pairs:

X	126	131	153	125	119	102	116	163
Y	120	126	152	129	102	105	100	175

test at a significance level not exceeding 0.10 the null hypothesis H_0: $M = 2$ against the alternative H_1: $M \neq 2$, where M is the median of the continuous and symmetric population of differences $D = X - Y$.

(d) Give the exact probability of a type I error in (c).

(e) Give the confidence interval corresponding to the test in (c).

5.5. Generate the sampling distributions of T^+ and T^- under the null hypothesis for a random sample of six unequal and nonzero observations.

5.6. Show by calculations from tables that the normal distribution provides reasonably accurate approximations to the critical values of one-sided tests for $\alpha = 0.01, 0.05$, and 0.10 when:

$N = 12$ for the sign test

$N = 15$ for the signed-rank test

5.7. A random sample of 10 observations is drawn from a normal population with mean μ and variance 1. Instead of a normal-theory test, the ordinary sign test is used for H_0: $\mu = 0$, H_1: $\mu > 0$, with rejection region $K \in R$ for $K \geqslant 8$.

(a) Plot the power curve using the exact distribution of K.

(b) Plot the power curve using the normal approximation to the distribution of K.

(c) Discuss how the power functions might help in the choice of an appropriate sample size for an experiment.

5.8. Prove that the Wilcoxon signed-rank statistic $T^+ - T^-$ based on a set of nonzero observations $X_1, X_2, \dots, X_N$ can be written symbolically in the form

$$\sum\sum_{1 \leqslant i \leqslant j \leqslant N} \operatorname{sgn}(X_i + X_j)$$

where

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

5.9. Let $D_1, D_2, \dots, D_N$ be a random sample of N nonzero observations from some continuous population which is symmetric with median zero. Define

$$|D_i| = \begin{cases} X_i & \text{if } D_i > 0 \\ Y_i & \text{if } D_i < 0 \end{cases}$$

Assume there are $m\,X$ values and $n\,Y$ values, where $m + n = N$ and the X and Y value are independent. Show that the signed-rank test statistic T^+ calculated for these D_i is equal to the sum of the ranks of the X observations in the combined ordered sample of mX's and nY's and also that $T^+ - T^-$ is the sum of the X ranks minus the sum of the Y ranks. This sum of the ranks of the X's is the test criterion for the Wilcoxon statistic in the two-sample problem to be discussed in Chapter 8. Show how T^+ might be used to test the hypothesis that the X and Y populations are identical.

5.10. Hoskin et al. (1986) investigated the change in fatal motor-vehicle accidents after the legal minimum drinking age was raised in 10 states. Their data were the ratios of the number of single-vehicle nighttime fatalities to the number of licensed drivers in the affected age group before and after the laws were changed to raise the drinking age, shown in Table 1. The researchers hypothesized that raising the minimum drinking age resulted in a reduced median fatality ratio. Investigate this hypothesis.

Table 1 Data for Problem 5.10

State	*Affected ages*	*Ratio before*	*Ratio after*
Florida	18	0.262	0.202
Georgia	18	0.295	0.227
Illinois	19–20	0.216	0.191
Iowa	18	0.287	0.209
Maine	18–19	0.277	0.299
Michigan	18–20	0.223	0.151
Montana	18	0.512	0.471
Nebraska	19	0.237	0.151
New Hampshire	18–19	0.348	0.336
Tennessee	18	0.342	0.307

5.11. The conclusion in Problem 5.10 was that the median difference (Before−After) was positive for the affected age group, but this does not imply that the reduction was the result of laws that raised the minimum legal drinking age. Other factors, counter measures, or advertising campaigns [like MADD (Mothers Against Drunk Drivers] may have affected the fatality ratios. In order to investigate further, these researchers compared the Before – After ratios for the affected age group with the corresponding difference ratios for the 25–29 age group, who were not affected by the law change, as shown in Table 2. Carry out an appropriate test and write a report of your conclusions.

Table 2 Data for Problem 5.11

State	*Affected age group*	25–29 *age group*
Florida	0.060	−0.025
Georgia	0.068	−0.023
Illinois	0.025	0.004
Iowa	0.078	−0.008
Maine	−0.022	0.061
Michigan	0.072	0.015
Montana	0.041	−0.035
Nebraska	0.086	−0.016
New Hampshire	0.012	−0.061
Tennessee	0.035	−0.051

5.12. Howard, Murphy, and Thomas (1986) reported a study designed to investigate whether computer anxiety changes between the beginning and end of a course on introduction to computers. The student subjects were given a test to measure computer anxiety at the beginning of the term and then again at the end of the 5-week summer course. High scores on this test indicate a high level of anxiety. For the data in Table 3 on 14 students, determine whether computer anxiety was reduced over the term.

Table 3 Data for Problem 5.12

Student	*Before*	*After*	*Student*	*Before*	*After*
A	20	20	H	34	19
B	21	18	I	28	13
C	23	10	J	20	21
D	26	16	K	29	12
E	32	11	L	22	15
F	27	20	M	30	14
G	38	20	N	25	17

5.13. Twenty-four students took both the midterm and the final exam in a writing course. Numerical grades were not given on the final, but each student was classified as either no change, improvement, or reduced level of performance compared with the midterm. Six showed improvement, 5 showed no change, and 13 had a reduced level of performance. Find the P value for an appropriate one-sided test.

5.14. Reducing high blood pressure by diet requires reduction of sodium intake, which usually requires switching from processed foods to their natural counterparts.

Listed below are the average sodium contents of five ordinary foods in processed form and natural form for equivalent quantities. Find a confidence interval estimate of the median difference (processed minus natural) with confidence coefficient at least 0.87 using *two different* procedures.

Natural food		*Processed food*	
Corn of the cob	2	Canned corn	251
Chicken	63	Fried chicken	1220
Ground Sirloin	60	All-beef frankfurter	461
Beans	3	Canned beans	300
Fresh tuna	40	Canned tuna	409

5.15. For the data in Problem 4.20, use both the sign test and the signed-rank test to investigate the research hypothesis that median earnings exceed 2.0.

5.16. In an experiment to measure the effect of mild intoxication on coordination, nine subjects were each given ethyl alcohol in an amount equivalent to 15.7 ml/m^2 of body surface and then asked to write a certain phrase as many times as they could in 1 min. The number of correctly written words was then counted and scaled such that a zero score represents the score a person not under the influence of alcohol would make, a positive score indicates increased writing speed and accuracy, and a negative score indicates decreased writing speed and accuracy. For the data below, find a confidence interval estimate of the median score at level nearest 0.95 using the procedure corresponding to the

(a) Sign test
(b) Wilcoxon signed-rank test where we assume symmetry

Subject	*Score*	*Subject*	*Score*
1	10	6	0
2	−8	7	−7
3	−6	8	5
4	−2	9	−8
5	15		

5.17. For the data in Example 4.3, test $H_0: M = 0.50$ against the alternative $H_1: M > 0.50$, using the

(a) Sign test
(b) Signed-rank test and assuming symmetry

5.18. For the data in Example 7.1, find a confidence interval estimate of the median difference Before minus After using the level nearest 0.90.

5.19. In a trial of two types of rain gauge, 69 of type A and 12 of type B were distributed at random over a small area. In a certain period 14 storms occurred, and the average amounts of rain recorded for each storm by the two types of gauge are as follows:

Another user claims to have found that the type B gauge gives consistently higher average readings than type A. Do these results substantiate such a conclusion? Investigate using two different nonparametric test procedures, by finding the P value from

Storm	*Type A*	*Type B*	*Storm*	*Type A*	*Type B*
1	1.38	1.42	8	2.63	2.69
2	9.69	10.37	9	2.44	2.68
3	0.39	0.39	10	0.56	0.53
4	1.42	1.46	11	0.69	0.72
5	0.54	0.55	12	0.71	0.72
6	5.94	6.15	13	0.95	0.90
7	0.59	0.61	14	0.55	0.52

(*a*) Tables of the exact distribution

(*b*) Large sample approximations to the exact distributions

(A total of four tests are to be performed.) Discuss briefly the advisability of using nonparametric versus parametric procedures for such an investigation and the relative merits of the two nonparametric tests used. Discuss assumptions in each case.

5.20. A manufacturer of suntan lotion is testing a new formula to see whether it provides more protection against sunburn than the old formula. The manufacturer chose 10 persons at random from among the company's employees, applied the two types of lotion to their backs, one type on each side, and exposed their backs to a controlled but intense amount of sun. Degree of sunburn was measured for each side of each subject, with the results shown below (higher numbers represent more severe sunburn).

(*a*) Test the null hypothesis that the difference (old – new) of degree of sunburn has median zero against the one-sided alternative that it is negative, assuming that the differences are symmetric. Does the new formula appear to be effective?

(*b*) Find a confidence interval for the median difference, assuming symmetry and with confidence coefficient near 0.90.

(*c*) Do (*a*) and (*b*) without assuming symmetry.

Subject	*Old formula*	*New formula*
1	41	37
2	42	39
3	48	31
4	38	39
5	38	34
6	45	47
7	21	19
8	28	30
9	29	25
10	14	8

5.21. Last year the elapsed time of long-distance telephone calls for a national retailer was skewed to the right with a median of 3 min 15 sec. The recession has reduced sales, but the company's treasurer claims that the median length of long-distance calls now is even greater than last year. A random sample of 5625 calls is selected from recent records and 2890 of them are found to last more than 3 min 15 sec. Is the treasurer's claim supported? Give the null and alternative hypotheses and the P value.

5.22. In order to test the effectiveness of a sales training program proposed by a firm of training specialists, a home furnishings company selects six sales representatives

Representative	*Sales before*	*Sales after*
1	90	97
2	83	80
3	105	110
4	97	93
5	110	123
6	78	84

at random to take the course. The data below are gross sales by these representatives before and after the course.

(*a*) State the null and alternative hypotheses and use the sign test to find a P value relevant to the question of whether the course is effective.

(*b*) Use the sign-test procedure at level nearest 0.90 to find a two-sided confidence-interval estimate of the median difference in sales (after – before). Give the exact level.

(*c*) Use the signed-rank test to do (*a*). What assumptions must you make?

(*d*) Use the signed-rank test procedure to do (*b*).

5.23. In a marketing research test, 15 adult males were asked to shave one side of their face with a brand A razor blade and the other side with a brand B razor blade and state their preferred blade. Twelve men preferred brand A. Find the P value for the alternative that the probability of preferring brand A is greater than 0.5.

5.24. Let X be a continuous random variable symmetrically distributed about θ. Show that the random variables $|X|$ and Z are independent, where

$$Z = \begin{cases} 1 & \text{if } X > \theta \\ 0 & \text{if } X \leqslant \theta \end{cases}$$

5.25. Using the result in Problem 5.24, show that for the Wilcoxon signed-rank test statistic T^+ discussed in Section 5.7, the $2N$ random variables $Z_1, r(|D_1|), Z_2, r(|D_2|), \ldots, Z_N\, r(|D_N|)$ are mutually independent under H_0.

5.26. Again consider the Wilcoxon signed-rank test discussed in Section 5.7. Show that under H_0 the distribution of the test statistic T^+ is the same as that of $W = \sum_{i=1}^{N} W_i$, where $W_1, W_2, \ldots, W_N$ are independent random variables with $P(W_i = 0) = P(W_i = i) = 0.5$, $i = 1, 2, \ldots, N$.

5.27. A study 5 years ago reported that the median amount of sleep by American adults is 7.5 hours out of 24 with a standard deviation of 1.5 hours and that 5% of the population sleep 6 or less hours while another 5% sleep 9 or more hours. A current sample of eight adults reported their average amounts of sleep per 24 hours as 7.2, 8.3, 5.6, 7.4, 7.8, 5.2, 9.1, and 5.8 hours. Use the most appropriate statistical procedures to determine whether American adults sleep less today than they did five years ago and justify your choice. You should at least test hypothesis concerning the quantiles of order 0.05, 0.50, and 0.95.

5.28. Find a confidence interval estimate of the median amount of sleep per 24 hours for the data in Problem 5.27 using confidence coefficient nearest 0.90.

5.29. Let $X_{(r)}$ denote the rth-order statistic of a random sample of size 5 from any continuous population and κ_p denote the pth quantile of this population. Find:

(*a*) $P(X_{(1)} < \kappa_{0.5} < X_{(5)})$
(*b*) $P(X_{(1)} < \kappa_{0.25} < X_{(3)})$
(*c*) $P(X_{(4)} < \kappa_{0.80} < X_{(5)})$

5.30. For order statistics of a random sample of size n from any continuous population F_X, show that the interval $(X_{(r)}, X_{(n-r+1)}, r < n/2)$, is a $100(1-\alpha)$ percent confidence-interval estimate for the median of F_X, where

$$1 - \alpha = 1 - 2n\binom{n-1}{r-1}\int_0^{0.5} x^{n-r}(1-x)^{r-1}\,dx$$

5.31. If $X_{(1)}$ and $X_{(\mathrm{n})}$ are the smallest and largest values, respectively, in a sample of size n from any continuous population F_X with median $\kappa_{0.50}$, find the smallest value of n such that:

(*a*) $P(X_{(1)} < \kappa_{0.50} < X_{(n)}) \geqslant 0.99$
(*b*) $P[F_X(X_{(n)}) - F_X(X_{(1)}) \geqslant 0.5] \geqslant 0.95$

5.32. Derive the sample size formula based on the normal approximation for the sign test against a two-sided alternative with approximate size α and power $1-\beta$.

5.33. Derive the sample size formula based on the normal approximation for the signed rank test against a two-sided alternative with approximate size α and power $1-\beta$.

6

The General Two-Sample Problem

6.1 INTRODUCTION

For the matched-pairs sign and signed-rank tests of Chapter 5 the data consisted of two samples, but each element in one sample was linked with a particular element of the other sample by some unit of association. This sampling situation can be described as a case of two dependent samples or alternatively as a single sample of pairs from a bivariate population. When the inferences to be drawn are related only to the population of differences of the paired observations, the first step in the analysis usually is to take the differences of the paired observations; this leaves only a single set of observations. Therefore, this type of data may be legitimately classified as a one-sample problem. In this chapter we shall be concerned with data consisting of two mutually independent random samples, i.e., random samples drawn independently from each of two populations. Not only are the elements within each sample independent, but also every element in the first sample is independent of every element in the second sample.

The universe consists of two populations, which we call the X and Y populations, with cumulative distribution functions denoted by F_X and F_Y, respectively. We have a random sample of size m drawn from the X population and another random sample of size n drawn independently from the Y population,

$$X_1, X_2, \ldots, X_m \qquad \text{and} \qquad Y_1, Y_2, \ldots, Y_n$$

Usually the hypothesis of interest in the two-sample problem is that the two samples are drawn from identical populations, i.e.,

$$H_0\colon F_Y(x) = F_X(x) \qquad \text{for all } x$$

If we are willing to make parametric model assumptions concerning the forms of the underlying populations and assume that the differences between the two populations occur only with respect to some parameters, such as the means or the variances, it is often possible to derive the so-called best test in a Neyman-Pearson framework. For example, if we assume that the populations are normally distributed, it is well known that the two-sample Student's t test for equality of means and the F test for equality of variances are respectively the best tests. The performances of these two tests are also well known. However, these and other classical tests may be sensitive to violations of the fundamental model assumptions inherent in the derivation and construction of these tests. Any conclusions reached using such tests are only as valid as the underlying assumptions made. If there is reason to suspect a violation of any of these postulates, or if sufficient information to judge their validity is not available, or if a completely general test of equality for unspecified distributions is desired, some nonparametric procedure is in order.

In practice, other assumptions are often made about the form of the underlying populations. One common assumption is called the *location model*, or the *shift model*. This model assumes that the X and Y populations are the same in all other respects except possibly for a shift in the (unknown) amount of say θ, or that

$$F_Y(x) = P(Y \leqslant x) = P(X \leqslant x - \theta) = F_X(x - \theta) \qquad \text{for all } x \text{ and } \theta \neq 0$$

This means that $X + \theta$ and Y have the same distribution or that X is distributed as $Y - \theta$. The Y population is then the same as the X population if $\theta = 0$, is shifted to the right if $\theta > 0$, and is shifted to the left if $\theta < 0$. Under the shift assumption, the populations have the same shape and the same variance, and the amount of the shift θ must be equal to the difference between the population means, $\mu_Y - \mu_X$, the population medians, $M_Y - M_X$, and in fact the difference between any two respective location parameters or quantiles of the same order.

Another assumption about the form of the underlying population is called the *scale model*, which assumes that the X and Y populations are the same except possibly for a positive scale factor θ which is not equal to one. The scale model can be written as

$$F_Y(x) = P(Y \leqslant x) = P(X \leqslant \theta) = F_X(\theta x) \qquad \text{for all } x \text{ and } \theta > 0, \theta \neq 1$$

This means that X/θ and Y have the same distribution for any positive θ or that X is distributed as θY. Also, the variance of X is θ^2 times the variance of Y and the mean of X is θ times the mean of Y.

A more general assumption about the form of the underlying populations is called a *location-scale model*. This model can be written as

$$P(Y - \mu_Y \leqslant x) = P(X - \mu_X \leqslant \theta)$$

which states the $(X - \mu_X)/\theta$ and $Y - \mu_Y$ are identically distributed (or similarly in terms of M_Y, M_X). Thus the location-scale model incorporates properties of both the location and the scale models. Now the means of $X - \mu_X$ and $Y - \mu_Y$ are both zero and the variance of $X - \mu_X$ is θ^2 times the variance of $Y - \mu_Y$.

Regardless of the model assumed, the general two-sample problem is perhaps the most frequently discussed problem in nonparametric statistics. The null hypothesis is almost always formulated as identical populations with the common distribution completely unspecified except for the assumption that it is a continuous distribution function. Thus under the null case, the two random samples can be considered a single random sample of size $N = m + n$ drawn from the common, continuous, but unspecified population. Then the combined ordered configuration of the m X and n Y random variables in the sample is one of the $\binom{m+n}{m}$ possible equally likely arrangements. For example, suppose we have two independent random samples, $m = 3$ X's and $n = 2$ Y's. Under the null hypothesis that the X's and the Y's are identically distributed, each of the $\binom{5}{2} = 10$ possible arrangements of the combined sample shown below is equally likely.

1. *XXXYY*	2. *XXYXY*	3. *YXYXX*	4. *XXYYX*	5. *XYXXY*
6. *XYXYX*	7. *YXXXY*	8. *YXXYX*	9. *XYYXX*	10. *YYXXX*

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

In practice, the sample pattern of arrangement of X's and Y's provides information about the type of difference which may exist in the populations. For instance, if the observed arrangement is that designated by either 1 or 10 in the above example, the X's and the Y's do not appear to be randomly mixed, suggesting a contradiction to the null hypothesis. Many statistical tests are based on some function of this combined arrangement. The type of function which is most appropriate depends on the type of difference one hopes to detect, which is indicated by the alternative hypothesis. An abundance of reasonable alternatives to H_0 may be considered, but the type easiest to analyze using distribution-free techniques states some functional relationship between the distributions. The most general two-sided alternative states simply

$$H_A\colon F_Y(x) \neq F_X(x) \qquad \text{for some } x$$

and a corresponding general one-sided alternative is

$$\begin{aligned} H_1\colon F_Y(x) &\geqslant F_X(x) \qquad \text{for all } x \\ F_Y(x) &> F_X(x) \qquad \text{for some } x \end{aligned}$$

In this latter case, we generally say that the random variable X is *stochastically larger* than the random variable Y. We can write this as $Y \overset{\text{ST}}{>} X$. Figures 1.1. and 1.2 are descriptive of the alternative that X is stochastically larger than Y, which includes as a subclass the more specific alternative $\mu_X > \mu_Y$. Some authors define $Y \overset{\text{ST}}{>} X$ to mean that $P(X > Y) > P(X < Y)$. (For the reverse inequality on F_X and F_Y, we say X is stochastically smaller than Y and write $X \overset{\text{ST}}{>} Y$).

If the particular alternative of interest is simply a difference in location, we use the location alternative or the location model

$$H_L\colon F_Y(x) = F_X(x - \theta) \qquad \text{for all } x \text{ and some } \theta \neq 0$$

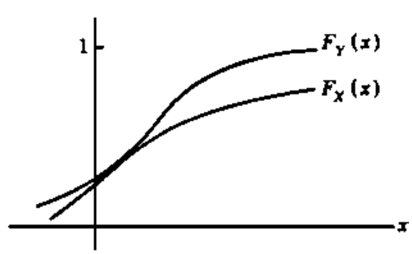


Fig. 1.1 X is stochastically larger than Y.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

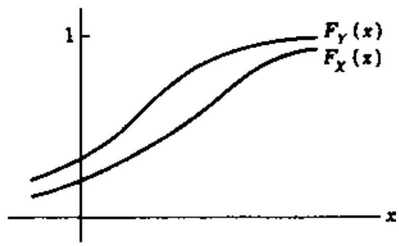


Fig. 1.2 X is stochastically larger than Y.

Under the location model, Y is distributed as $X + \theta$, so that Y is stochastically larger (smaller) than X if and only if $\theta > 0\,(\theta < 0)$. Similarly, if only a difference in scale is of interest, we use the scale alternative

$$H_S\colon F_Y(x) = F_X(\theta x) \qquad \text{for all } x \text{ and some } \theta \neq 1$$

Under the scale model, Y is distributed as X/θ, so that Y is stochastically larger (smaller) than X if and only if $\theta < 1\,(\theta > 1)$.

Although the three special alternatives H_1, H_L, and H_S are the most frequently encountered of all those included in the general class H_A, other types of relations may be considered. For example, the alternative $H_{\mathrm{LE}}\colon F_Y(x) = [F_X(x)]^k$, for some positive integer k and all x, called the *Lehmann alternative*, states that the Y random variables are distributed as the largest of k X variables. Under this alternative, Y is stochastically larger (smaller) than X if and only if $k > 1\,(k < 1)$.

The available statistical literature on the two-sample problem is quite extensive. A multitude of tests have been proposed for a wide variety of functional alternatives, but only a few of the best-known tests have been selected for inclusion in this book. The Wald-Wolfowitz runs test, the Kolmogorov-Smirnov two-sample test, the median test, the control median test, and the Mann-Whitney U test will be covered in this chapter. Chapters 7 and 8 are concerned with a specific class of tests particularly useful for the location and scale alternatives, respectively.

6.2 THE WALD-WOLFOWITZ RUNS TEST

Let the two sets of independent random variables $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be combined into a single ordered sequence from smallest to largest, keeping track of which observations correspond to the X sample and which to the Y. Assuming that their probability distributions are continuous, a unique ordering is always possible, since

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

theoretically ties do not exist. For example, with $m = 4$ and $n = 5$, the arrangement might be

$$X\ Y\ Y\ X\ X\ Y\ X\ Y\ Y$$

which indicates that in the pooled sample the smallest element was an X, the second smallest a Y, etc., and largest a Y. Under the null hypothesis of identical distributions

$$H_0\colon F_Y(x) = F_X(x) \qquad \text{for all } x$$

we expect the X and Y random variables to be well mixed in the ordered configuration, since the $m + n = N$ random variables constitute a single random sample of size N from the common population. With a run defined as in Chapter 3 as a sequence of identical letters preceded and followed by a different letter or no letter, the total number of runs in the ordered pooled sample is indicative of the degree of mixing. In our arrangement $X\ Y\ Y\ X\ X\ Y\ X\ Y\ Y$, the total number of runs is equal to 6 which shows a pretty good mixing of X's and Y's. A pattern of arrangement with too few runs would suggest that this group of N is not a single random sample but instead is composed of two samples from two distinguishable populations. For example, if the arrangement is $X\ X\ X\ X\ Y\ Y\ Y\ Y\ Y$ so that all the elements in the X sample are smaller than all of the elements in the Y sample, there would be only two runs. This particular configuration might indicate not only that the populations are not identical, but also that the X's are stochastically smaller than the Y's. However, the reverse ordering also contains only two runs, and therefore a test criterion based solely on the total number of runs cannot distinguish these two cases.

The runs test is appropriate primarily when the alternative is completely general and two-sided, as in

$$H_A\colon F_Y(x) \neq F_X(x) \qquad \text{for some } x$$

We define the random variable R as the total number of runs in the combined ordered arrangement of m X and n Y random variables. Since too few runs tend to discredit the null hypothesis when the alternative is H_A, the *Wald-Wolfowitz* (1940) *runs* test for significance level α generally has the rejection region in the lower tail as

$$R \leqslant c_\alpha$$

where c_α is chosen to be the largest integer satisfying

$$P(R \leqslant c_\alpha \mid H_0) \leqslant \alpha$$

The P value for the runs test is then given by

$$P(R \leqslant R_O \mid H_0)$$

where R_O is the observed value of the runs test statistic R.

Since the X and Y observations are two types of objects arranged in a completely random sequence if H_0 is true, the null probability distribution of R is exactly the same as was found in Chapter 3, for the runs test for randomness. The distribution is given in Theorem 2.2 of Section 3.2 with n_1 and n_2 replaced by m and n, respectively, assuming the X's are called type 1 objects and Y's are the type 2 objects. The other properties of R discussed in that section, including the moments and asymptotic null distribution, are also unchanged. The only difference here is that the appropriate critical region for the alternative of different populations is too few runs. The null distribution of R is given in Table D of the Appendix with $n_1 = m$ and $n_2 = n$ for $m \leqslant n$. The normal approximation described in Section 3.2 is used for larger sample sizes. A numerical example of this test is given below.

Example 2.1 It is easy to show that the distribution of a standardized chi-square variable with large degrees of freedom can be approximated by the standard normal distribution. This example provides an investigation of the agreement between these two distributions for moderate degrees of freedom. Two mutually independent random samples, each of size 8, were generated, one from the standard normal distribution and one from the chi-square distribution with $\nu = 18$ degrees of freedom. The resulting data are as follows:

Normal	−1.91	−1.22	−0.96	−0.72	0.14	0.82	1.45	1.86
Chi square	4.90	7.25	8.04	14.10	18.30	21.21	23.10	28.12

Solution Before testing the null hypothesis of equal distributions, the chi-square sample data must be standardized by subtracting the mean $\nu = 18$ and dividing by the standard deviation $\sqrt{2\nu} = \sqrt{36} = 6$. The transformed chi-square data are, respectively,

$$-2.18 \quad -1.79 \quad -1.66 \quad -0.65 \quad 0.05 \quad 0.54 \quad 0.85 \quad 1.69$$

We pool the normal data and these transformed data into a single array, ordering them from smallest to largest, underlining the transformed chi-square data, as

$$-\underline{2.18},\ -1.91,\ -\underline{1.79},\ \underline{1.66}, -1.22, -0.96, -0.72, -\underline{0.65},\ \underline{0.05},$$
$$0.14,\ \underline{0.54},\ 0.82,\ \underline{0.85},\ 1.45,\ \underline{1.69},\ 1.86$$

Let X denote the standardized chi-square sample data and let Y denote the normal sample data. For the solution using the Wald-Wolfowitz runs test, we simply count the number of runs in the ordered combined configuration $X, Y, X, X, Y, Y, Y, X, X, Y, X, Y, X, Y, X, Y$ as $R = 12$. Table D shows that the P value, the left-tail probability with $R = 12$ for $m = 8, n = 8$, exceeds 0.5, and therefore we do not reject the null hypothesis of equal distributions. Using (2.11) of Section 3.2 with a continuity correction, we get $Z = 1.81$ with $P = 0.9649$ and $z = 1.55$ or $P = 0.9394$ without a continuity correction.

The STATXACT solution to Example 2.1 using the runs test is shown below. Note that the exact P value is 0.9683. This can be verified from Table D since $P(R \leqslant 12) = 1 - P(R \geqslant 13) = 0.968$. Note that their asymptotic P value is not the same as ours using (2.11) of Section 3.2.

```
******************************************
      STATXACT SOLUTION TO EXAMPLE 2.1
******************************************

WALD WOLFOWITZ RUNS TEST

Summary of Exact distribution of WALD WOLFOWITZ RUNS TEST
statistic

          Min           Max       Observed
        2.000         16.00          12.00

Asymptotic p-value :
    Pr { Test Statistic .LE.        12.00 } =      0.9021

Exact p-value:
    Pr {  Test Statistic .LE.       12.00 } =      0.9683
    Pr {  Test Statistic .EQ.       12.00 } =      0.0685
```

THE PROBLEM OF TIES

Ideally, no ties should occur because of the assumption of continuous populations. Ties do not present a problem in counting the number of runs unless the tie is across samples; i.e., two or more observations from different samples have exactly the same magnitude. For a conservative test, we can break all ties in all possible ways and compute the total number of runs for each resolution of all ties. The actual R used as the value of the test statistic is the largest computed value, since that is the one least likely to lead to rejection of H_0. For each group of ties across samples, where there are s x's and t y's of equal magnitude for some $s \geqslant 1, t \geqslant 1$, there are $\binom{s+t}{s}$ ways to break the ties. Thus if there are k groups of ties, the total number of values of R to be computed is the product $\prod_{i=1}^{k} \binom{s_i + t_i}{s_i}$.

DISCUSSION

The Wald-Wolfowitz runs test is extremely general and is consistent against all types of differences in populations [Wald and Wolfowitz (1940)]. The very generality of the test weakens its performance against specific alternatives. Asymptotic power can be evaluated using the normal distribution with appropriate moments under the alternative, which are given in Wolfowitz (1949). Since power, whether exact or asymptotic, can be calculated only for completely specified alternatives, numerical power comparisons should not be the only criteria for this test. Its primary usefulness is in preliminary analyses of data when no particular form of alternative is yet formulated. Then, if the hypothesis is rejected, further studies can be made with other tests in an attempt to classify the type of difference between populations.

6.3 THE KOLMOGOROV-SMIRNOV TWO-SAMPLE TEST

The Kolmogorov-Smirnov statistic is another one-sample test that can be adapted to the two-sample problem. Recall from Chapter 4 that as a goodness-of-fit criterion, this test compared the empirical distribution function of a random sample with a hypothesized cumulative distribution. In the two-sample case, the comparison is made between the empirical distribution functions of the two samples.

The order statistics corresponding to two random samples of size m and n from continuous populations F_X and F_Y, are

$$X_{(1)}, X_{(2)}, \ldots, X_{(m)} \qquad \text{and} \qquad Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$$

Their respective empirical distribution functions, denoted by $S_m(x)$ and $S_n(x)$, are defined as before:

$$S_m(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ k/m & \text{if } X_{(k)} \leqslant x < X_{(k+1)} \qquad \text{for } k = 1, 2, \ldots, m-1 \\ 1 & \text{if } x \geqslant X_{(m)} \end{cases}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and

$$S_n(x) = \begin{cases} 0 & \text{if } x < Y_{(1)} \\ k/n & \text{if } Y_{(k)} \leqslant x < Y_{(k+1)} \qquad \text{for } k = 1, 2, \ldots, n-1 \\ 1 & \text{if } x \geqslant Y_{(n)} \end{cases}$$

In a combined ordered arrangement of the $m + n$ sample observations, $S_m(x)$ and $S_n(x)$ are the respective proportions of X and Y observations which do not exceed the specified value x.

If the null hypothesis

$$H_0\colon F_Y(x) = F_X(x) \qquad \text{for all } x$$

is true, the population distributions are identical and we have two samples from the same population. The empirical distribution functions for the X and Y samples are reasonable estimates of their respective population cdf. Therefore, allowing for sampling variation, there should be reasonable agreement between the two empirical distributions if indeed H_0 is true; otherwise the data suggest that H_0 is not true and therefore should be rejected. This is the intuitive logic behind most two-sample tests, and the problem is to define what is a reasonable agreement between the two empirical cdf's. In other words, how close do the two empirical cdf's have to be so that they could be viewed as not significantly different, taking account of the sampling variability. Note that this approach necessarily requires a definition of closeness. The two-sided *Kolmogorov-Smirnov two-sample test* criterion, denoted by $D_{m,n}$, is based on the maximum absolute difference between the two empirical distributions

$$D_{m,n} = \max_x |S_m(x) - S_n(x)|$$

Since here only the magnitudes, and not the directions, of the deviations are considered, $D_{m,n}$ is appropriate for a general two-sided alternative

$$H_A\colon F_Y(x) \neq F_X(x) \qquad \text{for some } x$$

and the rejection region is in the upper tail, defined by

$$D_{m,n} \geqslant c_\alpha$$

where

$$P(D_{m,n} \geqslant c_\alpha \mid H_0) \leqslant \alpha$$

Because of the Gilvenko-Cantelli theorem (Theorem 3.2 of Section 2.3), the test is consistent for this alternative. The P value is

$$P(D_{m,n} \geqslant D_O \mid H_0)$$

where D_O is the observed value of the two-sample K-S test statistic. As with the one-sample Kolmogorov-Smirnov statistic, $D_{m,n}$ is completely distribution free for any continuous common population distribution since order is preserved under a monotone transformation. That is, if we let $z = F(x)$ for the common continuous cdf F, we have $S_m(z) = S_m(x)$ and $S_n(z) = S_n(x)$, where the random variable Z, corresponding to z, has the uniform distribution on the unit interval.

In order to implement the test, the exact cumulative null distribution of $mnD_{m,n}$ is given in Table I in the Appendix for $2 \leqslant m \leqslant n \leqslant 12$ or $m + n \leqslant 16$, whichever occurs first. Selected quantiles of $mnD_{m,n}$ are also given for $m = n$ between 9 and 20, along with the large sample approximation.

The derivation of the exact null probability distribution of $D_{m,n}$ is usually attributed to the Russian School, particularly Gnedenko (1954) and Korolyuk (1961), but the papers by Massey (1951b, 1952) are also important. Several methods of calculation are possible, generally involving recursive formulas. Drion (1952) derived a closed expression for exact probabilities in the case $m = n$ by applying random-walk techniques. Several approaches are summarized in Hodges (1958). One of these methods, which is particularly useful for small sample sizes, will be presented here as an aid to understanding.

To compute $P(D_{m,n} \geqslant d \mid H_0)$, where d is the observed value of $\max_x |S_m(x) - S_n(x)|$, we first arrange the combined sample of $m + n$ observations in increasing order of magnitude. The arrangement can be depicted graphically on a Cartesian coordinate system by a path which starts at the origin and moves one step to the right for an x observation and one step up for a y observation, ending at (m,n). For example, the sample arrangement *xyyxxyy* is represented in Figure 3.1. The observed values of $mS_m(x)$ and $nS_n(x)$ are, respectively, the coordinates of all points (u,v) on the path where u and v are integers. The number d is the largest of the differences $|u/m - v/n| = |nu - mv|/mn$. If a line is drawn connecting the points $(0,0)$ and (m,n) on this graph, the equation of the line is $nx - my = 0$ and the vertical distance from any point (u,v) on the path to this line is $|v - nu/m|$. Therefore, nd for the observed sample is the distance from the diagonal line. In Figure 3.1 the farthest point is labeled Q, and the value of d is 2/4.

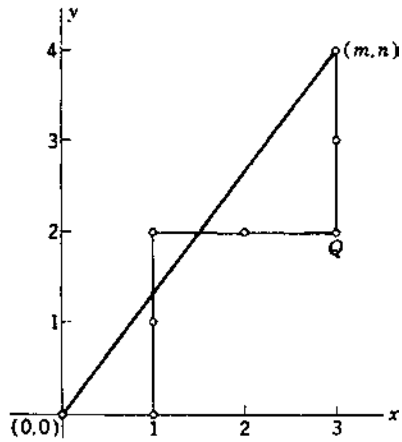


Fig. 3.1 Path of *xyyxxyy*.

The total number of arrangements of m X and n Y random variables is $\binom{m+n}{m}$, and under H_0 each of the corresponding paths is equally likely. The probability of an observed value of $D_{m,n}$ not less than d then is the number of paths which have points at a distance from the diagonal not less than nd, divided by $\binom{m+n}{m}$.

In order to count this number, we draw another figure of the same dimension as before and mark off two lines at vertical distance nd from the diagonal, as in Figure 3.2. Denote by $A(m,n)$ the number of paths from (0,0) to (m,n) which lie entirely *within* (not on) these boundary lines. Then the desired probability is

$$P(D_{m,n} \geqslant d \,|\, H_0) = 1 - P(D_{m,n} < d \,|\, H_0) = 1 - \frac{A(m,n)}{\binom{m+n}{m}}$$

$A(m,n)$ can easily be counted in the manner indicated in Figure 3.2. The number $A(u,v)$ at any intersection (u,v) clearly statisfies the recursion relation

$$A(u,v) = A(u-1,v) + A(u,v-1)$$

with boundary conditions

$$A(0,v) = A(u,0) = 1$$

Thus $A(u,v)$ is the sum of the numbers at the intersections where the previous point on the path could have been while still within the

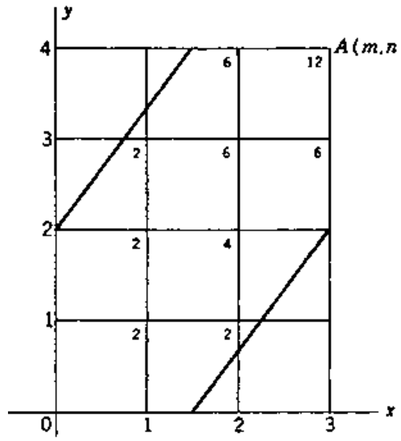


Fig. 3.2 Evaluation of $A(u, v)$ for *xyyxxyy*.

boundaries. This procedure is shown in Figure 3.2 for the arrangement *xyyxxyy*, where $nd = 2$. Since here $A(3,4) = 12$, we have

$$P(D_{3,4} \geqslant 0.5) = 1 - \frac{12}{\binom{7}{4}} = \frac{23}{35} = 0.65714$$

For the asymptotic null distribution, that is, m and n approach infinity in such a way that m/n remains constant, Smirnov (1939) proved the result

$$\lim_{m,n\to\infty} P\left(\sqrt{\frac{mn}{m+n}}\, D_{m,n} \leqslant d\right) = L(d)$$

where

$$L(d) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} e^{-2i^2 d^2}$$

Note that the asymptotic distribution of $\sqrt{mn/(m+n)}D_{m,n}$ is exactly the same as the asymptotic distribution of $\sqrt{N}D_N$ in Theorem 3.3 of Section 4.3. This is not surprising, since we know from the Glivenko-Cantelli theorem that as $n \to \infty, S_n(x)$ converges to $F_Y(x)$, which can be relabeled $F_X(x)$ as in the theorem. Then the only difference here is in the normalizing factor $\sqrt{mn/(m+n)}$, which replaces $\sqrt{N}$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

ONE-SIDED ALTERNATIVES

A one-sided two-sample maximum-unidirectional-deviation test can also be defined, based on the statistic

$$D^+_{m,n} = \max_x [S_m(x) - S_n(x)]$$

For an alternative that the X random variables are stochastically smaller than the Y's,

$$\begin{aligned} H_1\colon F_Y(x) &\leqslant F_X(x) \qquad \text{for all } x \\ F_Y(x) &< F_X(x) \qquad \text{for some } x \end{aligned}$$

the rejection region should be

$$D^+_{m,n} \geqslant c_\alpha$$

The one-sided test based on $D^+_{m,n}$ is also distribution free and consistent against the alternative H_1. Since either sample may be labeled the X sample, it is not necessary to define another one-sided statistic for the alternative that X is stochastically larger than Y. The entries in Table I in the Appendix can also be used for a one-sided two-sample Kolmogorov-Smirnov statistic since the probabilities in the tails of this distribution are closely approximated using one-half of the corresponding tail probabilities on the two-sided, two-sample Kolmogorov-Smirnov statistic.

The graphic method described for $D_{m,n}$ can be applied here to calculate $P(D^+_{m,n} \geqslant d)$. The point Q^+, corresponding to Q, would be the point farthest *below* the diagonal line, and $A(m,n)$ is the number of paths lying entirely *above* the lower boundary line (see Problem 6.1). Tables of the null distribution of $D^+_{m,n}$ are available in Goodman (1954) for $m = n$.

As with the two-sided statistic, the asymptotic distribution of $\sqrt{mn/(m+n)}D^+_{m,n}$ is equivalent to the asymptotic distribution of $\sqrt{N}D^+_N$, which was given in Theorem 3.5 of Section 4.3 as

$$\lim_{m,n\to\infty} P\left(\sqrt{\frac{mn}{m+n}}\, D^+_{m,n} \leqslant d\right) = 1 - e^{-2d^2}$$

TIES

Ties within and across samples can be handled by considering only the r distinct ordered observations in the combined sample as values of x in computing $S_m(x)$ and $S_n(x)$ for $r \leqslant m$ and $r \leqslant n$. Then we find the empirical cdf for each different x and their differences at these observations and calculate the statistic in the usual way.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

DISCUSSION

The Kolmogorov-Smirnov tests are very easy to apply, using the exact distribution for any m and n within the range of the available tables and using the asymptotic distribution for larger sample sizes. They are useful mainly for the general alternatives H_A and H_1, since the test statistic is sensitive to all types of differences between the cumulative distribution functions. Their primary application then should be for preliminary studies of data, as was the runs test. Gideon and Mueller (1978) give a simple method for calculating $D_{m,n}$ and Pirie (1979) extends this method to samples with ties. The Kolmogorov-Smirnov tests are more powerful than the runs tests when compared against the Lehmann (1953) type of nonparametric alternatives for large sample sizes. The large-sample performance of the Kolmogorov-Smirnov tests against specific location or scale alternatives varies considerably according to the population sampled. Capon (1965) has made a study of these properties. Goodman (1954) has shown that when applied to data from discrete distributions, these tests are conservative.

APPLICATIONS

Application of the Kolmogorov-Smirnov two-sample general test is illustrated below with the data from Example 2.1.

Example 3.1 To carry out the Kolmogorov-Smirnov two-sample test against the two-sided alternative, we calculate the two empirical distribution functions and their differences, as shown in Table 3.1. Note that the first column shows the combined (pooled) ordered sample. This is labeled t to avoid notational confusion. The maximum of the last column is $D_{m,n} = 2/8$ so that $mnD_{m,n} = 16$. Table I for $m = n = 8$, shows that $P(64D_{8,8} \geqslant 32 \mid H_0) = 0.283$, so the required P value, $P(64D_{8,8} \geqslant 16 \mid H_0)$, must be greater than 0.283. Thus, we do not reject the null hypothesis of identical distributions.

For the one-sided alternative, $D^+_{m,n} = 2/8$ and so the P value is at least $(0.283)/2 = 0.142$. Thus there is not sufficient evidence to reject H_0 against the one-sided alternative that the X's are stochastically smaller than the Y's.

The STATXACT solution to Example 3.1 using the Kolmogorov-Smirnov test is shown below. Note that for the two-sided alternative, the exact and the asymptotic P values are shown to be 0.9801 and 0.9639, respectively, both strongly suggesting that there is no significant evidence against the null hypothesis in these data. The

Table 3.1 Calculation of $D_{m,n}$ for Example 3.1

t	$\# X \leqslant t$	$S_m(t)$	$\# Y \leqslant t$	$S_n(t)$	$S_m(t) - S_n(t)$	$\lvert S_m(t) - S_n(t) \rvert$
−2.18	1	1/8	0	0	1/8	1/8
−1.91	1	1/8	1	1/8	0	0
−1.79	2	2/8	1	1/8	1/8	1/8
−1.66	3	3/8	1	1/8	2/8	2/8
−1.22	3	3/8	2	2/8	1/8	1/8
−0.96	3	3/8	3	3/8	0	0
−0.72	3	3/8	4	4/8	−1/8	1/8
−0.65	4	4/8	4	4/8	0	0
0.05	5	5/8	4	4/8	1/8	1/8
0.14	5	5/8	5	5/8	0	0
0.54	6	6/8	5	5/8	1/8	1/8
0.82	6	6/8	6	6/8	0	0
0.85	7	7/8	6	6/8	1/8	1/8
1.45	7	7/8	7	7/8	0	0
1.69	8	8/8	7	7/8	1/8	1/8
1.86	8	8/8	8	8/8	0	0

exact two-sided P value is a Monte Carlo estimate; the algorithm is described in Hilton, Mehta and Patel (1994). The asymptotic two-sided P value is calculated using the Smirnov approximation, keeping only the first few terms. The exact one-sided P value is calculated from the permutation distribution of $D^+_{m,n}$. The reader is referred to the STATXACT user manual for details.

```
          ******************************************
          STATXACT SOLUTION TO EXAMPLE 3.1: K-S TEST
          ******************************************

KOLMOGOROV-SMIRNOV TWO-SAMPLE TEST
POP_1 (F1) :        1         POP_2 (F2) :        2

      Number of Observations:
             POP_1 = 8
             POP_2 = 8

                        |F1 - F2|          F1 - F2              F2 - F1
                                    (POP_1 is larger) (POP_2 is larger)
Observed Statistic   0.2500             0.1250               0.2500
Asymptotic p-value   0.9639             0.8825               0.6065
Exact p-value        0.9801             0.8889               0.6222
Exact Point Prob.    0.3200             0.2667               0.2828
```

6.4 THE MEDIAN TEST

In order to test the null hypothesis of identical populations with two independent samples, the Kolmogorov-Smirnov two-sample test compares the proportions of observations from each sample which do not exceed some number x for all real numbers x. The test criterion was the maximum difference (absolute or unidirectional) between the two empirical distributions, which are defined for all x. Suppose that instead of using all possible differences, we choose some arbitrary but specific number δ and compare only the proportions of observations from each sample which are strictly less than δ. As before, the two independent samples are denoted by

$$X_1, X_2, \ldots, X_m \qquad \text{and} \qquad Y_1, Y_2, \ldots, Y_n$$

Each of the $m + n = N$ observations is to be classified according to whether it is less than δ or not. Let U and V denote the respective numbers of X and Y observations less than δ. Since the random variables in each sample have been dichotomized, U and V both follow the binomial probability distribution with parameters

$$p_X = P(X < \delta) \qquad \text{and} \qquad p_Y = P(Y < \delta)$$

and numbers of trials m and n, respectively. For two independent samples, the joint distribution of U and V then is

$$f_{U,V}(u,v) = \binom{m}{u}\binom{n}{v} p_X^u p_Y^v (1 - p_X)^{m-u}(1 - p_Y)^{n-v} \qquad u = 0, 1, \ldots, m \text{ and } v = 0, 1, \ldots, n \tag{4.1}$$

The random variables U/m and V/n are unbiased point estimates of the parameters p_X and p_Y, respectively. The difference $U/m - V/n$ then is appropriate for testing the null hypothesis

$$H_0\colon p_X - p_Y = 0$$

The exact null probability distribution of $U/m - V/n$ can easily be found from (4.1), and for m and n large its distribution can be approximated by the normal. The test statistic in either case depends on the common value $p = p_X = p_Y$, but the test can be performed by replacing p by its unbiased estimate $(u + v)/(m + n)$. Otherwise there is no difficulty in constructing a test (although approximate) based on the criterion of difference of proportions of observations less than δ. This is essentially a modified sign test for two independent samples,

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

with the hypothesis that δ is the pth quantile point in both populations, where p is unspecified but estimated from the data.

This test will not be pursued here since it is approximate and is not always appropriate to the general two-sample problem, where we are primarily interested in the hypothesis of identical populations. If the two populations are the same, the pth quantile points are equal for every value of p. However, two populations may be quite disparate even though some particular quantile points are equal. The value of δ, which is supposedly chosen without knowledge of the observations, then affects the sensitivity of the test criterion. If δ is chosen too small or too large, both U and V will have too small a range to be reliable. We cannot hope to have reasonable power for the general test without a judicious choice of δ. A test where the experimenter chooses a particular value of p (rather than δ), preferably a central value, would be more appropriate for our general hypothesis, especially if the type of difference one hopes to detect is primarily in location. In other words, we would rather control the position of δ, regardless of its actual value, but p and δ are hopelessly interrelated in the common population.

When the populations are assumed identical but unspecified, we cannot choose p and then determine the corresponding δ. Yet δ must be known at least positionally to classify each sample observation as less than δ or not. Therefore, suppose we decide to control the position of δ relative to the magnitudes of the *sample* observations. If the quantity $U + V$ is fixed by the experimenter prior to sampling, p is to some extent controlled since $(u + v)/(m + n)$ is an estimate of the common p. If p denotes the probability that any observation is less than δ, the probability distribution of $T = U + V$ is

$$f_T(t) = \binom{m+n}{t} p^t (1-p)^{m+n-t} \qquad t = 0, 1, \ldots, m+n \tag{4.2}$$

The conditional distribution of U given $T = t$ is (4.1) divided by (4.2). In the null case where $p_X = p_Y = p$, the result is simply

$$f_{U|T}(u|t) = \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}} \qquad u = \max(0,\, t-n), 1, \ldots, \min(m, t) \tag{4.3}$$

which is the hypergeometric probability distribution. This result could also have been argued directly as follows. Each of the $m + n$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

observations is dichotomized according to whether it is less than δ or not. Among all the observations, if $p_X = p_Y = p$, every one of the $\binom{m+n}{t}$ sets of t numbers is equally likely to comprise the less-than-δ group.

The number of sets that have exactly u from the X sample is $\binom{m}{u}\binom{n}{t-u}$. Since U/m is an estimate of p_X, if the hypothesis $p_X = p_Y = p$ is true, u/m should be close to $t/(m+n)$. A test criterion can then be found using the conditional distribution of U in (4.3) for any chosen t.

So far nothing has been said about the value of δ, since once t is chosen, δ really need not be specified to perform the test. Any number greater than the tth and not greater than the $(t+1)$st order statistic in the combined ordered sample will yield the same value of u. In practice, the experimenter would probably rather choose the fraction $t/(m+n)$ in order to control the value of p. Suppose we decide that if the populations differ at all, it is only in location. Then a reasonable choice of $t/(m+n)$ is 0.5. But $N = m+n$ may be odd or even, while t must be an integer. To eliminate inconsistencies in application, δ can be defined as the $[(N+1)/2]$nd order statistic if N is odd, and any number between the $(N/2)$nd and $[(N+2)/2]$nd order statistics for N even. Then a unique value of u is obtained for any set of N observations, and δ is actually defined to be the median of the combined samples. The probability distribution of U is given in (4.3), where $t = N/2$ for N even and $t = (N-1)/2$ for N odd. The test based on U, the number of observations from the X sample which are less than the combined sample median, is called the *median test*. It is attributed mainly to Brown and Mood (1948, 1951), Mood (1950), and Westenberg (1948) and is often referred to as Mood's median test or the joint median test.

The fact that δ cannot be determined before the samples are taken may be disturbing, since it implies that δ should be treated as a random variable. In deriving (4.3) we treated δ as a constant, but the same result is obtained for δ defined as the sample median value. Denote the combined sample median by the random variable Z and the cdf's of the X and Y populations by F_X and F_Y, respectively, and assume that N is odd. The median Z can be either an X or a Y random variable, and these possibilities are mutually exclusive. The joint density of U and Z for t observations less than the sample median where $t = (N-1)/2$ is the limit, as Δz approaches zero, of the sum of the probabilities that (1) the X's are divided into three classifications, u less than z, one between z and $z + \Delta z$ and the remainder greater than $z + \Delta z$, and the Y's are divided such that $t - u$ are less than z, and (2) exactly u X's are less than z, and the Y's are divided such that $t-u$ are less than z, one is between z and $z+\Delta z$, and the remainder are greater than $z+\Delta z$. The result then is

$$
\begin{aligned}
f_{U,Z}(u,z) = & \binom{m}{u, 1, m-1-u} [F_X(z)]^u f_X(z) \\
& \times [1-F_X(z)]^{m-1-u} \binom{n}{t-u} [F_Y(z)]^{t-u} [1-F_Y(z)]^{n-t+u} \\
& + \binom{m}{u} [F_X(z)]^u [1-F_X(z)]^{m-u} \binom{n}{t-u, 1, n-t+u-1}^{t-u} \\
& \times [F_Y(z)]^{t-u} f_Y(z) [1-F_Y(z)]^{n-t+u-1}
\end{aligned}
$$

The marginal density of U is obtained by integrating the above expression over all z, and if $F_X(z) = F_Y(z)$ for all z, the result is

$$
\begin{aligned}
f_U(u) &= \left[m\binom{m-1}{u}\binom{n}{t-u} + n\binom{m}{u}\binom{n-1}{t-u} \right] \\
&\quad \times \int_{-\infty}^{\infty} [F(z)]^t [1-F(z)]^{m+n-t-1} f(z)\, dz \\
&= \binom{m}{u}\binom{n}{t-u} [(m-u) + (n-t+u)] B(t+1, m+n-t) \\
&= \binom{m}{u}\binom{n}{t-u} \frac{t!(m+n-t)!}{(m+n)!}
\end{aligned}
$$

which agrees with the expression in (4.3).

Because of this result, we might say then that before sampling, i.e., before the value of δ is determined, the median test statistic is appropriate for the general hypothesis of identical populations, and after the samples are obtained, the hypothesis tested is that δ is the pth quantile value in both populations, where p is a number close to 0.5. The null distributions of the test statistic are the same for both hypotheses, however.

Even though the foregoing discussion may imply that the median test has some statistical and philosophical limitations in conception, it is well known and accepted within the context of the general two-sample problem. The procedure for two independent samples of measurements is to arrange the combined samples in increasing order of magnitude and determine the sample median δ, the observation with rank $(N+1)/2$ if N is odd and any number between the observations

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

with rank $N/2$ and $(N+2)/2$ if N is even. A total of t observations are then less than δ, where $t = (N-1)/2$ or $N/2$ according as N is odd or even. Let U denote the number of X observations less than δ. If the two samples are drawn from identical continuous populations, the probability distribution of U for fixed t is

$$f_U(u) = \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}}$$
$$u = \max(0, t-n), \ldots, \min(m, t), \quad t = [N/2] \tag{4.4}$$

where $[x]$ denotes the largest integer not exceeding the value x. If the null hypothesis is true, then $P(X < \delta) = P(Y < \delta)$ for all δ, and in particular the two populations have a common median, which is estimated by δ.

Since U/m is an estimate of $P(X < \delta)$, which is approximately one-half under H_0, a test based on the value of U will be most sensitive to differences in location. If U is much larger than $m/2$, most of the X values are less than most of the Y values. This lends credence to the relation $P(X < \delta) > P(Y < \delta)$, that the X's are stochastically smaller than the Y's, so that the median of the X population is smaller than the median of the Y population, or that $\theta > 0$. If U is too small relative to $m/2$, the opposite conclusion is implied. The appropriate rejection regions and P values for nominal significance level α then are as follows:

Alternative	*Rejection region*	*P value*
$Y \overset{ST}{>} X$, $\theta > 0$ or $M_Y > M_X$	$U \geqslant c'_\alpha$	$P(U \geqslant U_O)$
$Y \overset{ST}{<} X$, $\theta < 0$ or $M_Y < M_X$	$U \leqslant c_\alpha$	$P(U \leqslant U_O)$
$\theta \neq 0$ or $M_X \neq M_Y$	$U \leqslant c$ or $U \geqslant c'$	2 (smaller of the above)

where c_α and c'_α are, respectively, the largest and smallest integers such that $P(U \leqslant c_\alpha \mid H_0) \leqslant \alpha$ and $P(U \geqslant c'_\alpha \mid H_0) \leqslant \alpha$, c and c' are two integers $c < c'$ such that

$$P(U \leqslant c \mid H_0) + P(U \geqslant c' \mid H_o) \leqslant \alpha$$

and U_O is the observed value of the median test statistic U.

The critical values c and c' can easily be found from (4.4) or from tables of the hypergeometric distribution [Lieberman and Owen (1961)] or by using tables of the binomial coefficients. If N is even, we choose $c'_\alpha = m - c_\alpha$. Since the distribution in (4.4) is not symmetric for $m \neq n$ if N is odd, the choice of an optimum rejection region for a two-sided test is not clear for this case. It could be chosen such that α is divided equally or that the range of u is symmetric, or neither.

If m and n are so large that calculation or use of tables to find critical values is not feasible, a normal approximation to the hypergeometric distribution can be used. Using formulas for the mean and the variance of the hypergeometric distribution (given in Chapter 1) and the distribution in (4.4), the mean and variance of U are easily found to be

$$E(U \mid t) = \frac{mt}{N} \qquad \text{var}(U \mid t) = \frac{mnt(N-t)}{N^2(N-1)} \tag{4.5}$$

If m and n approach infinity in such a way that m/n remains constant, this hypergeometric distribution approaches the binomial distribution for t trials with parameter m/N, which in turn approaches the normal distribution. For N large, the variance of U in (4.5) is approximately

$$\text{var}(U \mid t) = \frac{mnt(N-t)}{N^3}$$

and thus the asymptotic distribution of

$$Z = \frac{U - mt/N}{[mnt(N-t)/N^3]^{1/2}} \tag{4.6}$$

is approximately standard normal. A continuity correction of 0.5 may be used to improve the approximation. For example, when the alternative is $\theta < 0$ (or $M_Y < M_X$), the approximate P value with a continuity correction is given by

$$\Phi\left(\frac{U_O + 0.5 - mt/N}{\sqrt{mnt(N-t)/N^3}}\right) \tag{4.7}$$

It is interesting to note that a test based on the statistic Z in (4.6) is equivalent to the usual normal-theory test for the difference between two independent proportions found in most statistics books.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

This can be shown by algebraic manipulation of (4.6) with $t = u + v$ as follows

$$z = \frac{Nu - mt}{\sqrt{mnt(N-t)/N}} = \frac{nu - m(t-u)}{\sqrt{mnN(t/N)(1-t/N)}}$$
$$= \frac{u/m - v/n}{\sqrt{[(u+v)/N][1-(u+v)/N]N/mn}}$$
$$= \frac{u/m - v/n}{\sqrt{\hat{p}(1-\hat{p})(1/m+1/n)}}$$

If a success is defined as an observation being less than δ, u/m and v/n are the observed sample proportions of successes, and $\hat{p} = (u+v)/N$ is the best sample estimate of the common proportion. This then is the same approximate test statistic that was described at the beginning of this section for large samples, except that here $u + v = t$, a constant which is fixed by the choice of δ as the sample median.

The presence of ties either within or across samples presents no problem for the median test except in two particular cases. If N is odd and more than one observation is equal to the sample median, or if N is even and the $(N/2)$nd- and $[(N+2)/2]$nd-order statistics are equal, t cannot be defined as before unless the ties are broken. The conservative approach is recommended, where the ties are broken in all possible ways and the value of u chosen for decision is the one which is least likely to lead to rejection of H_0.

APPLICATIONS

Example 4.1 The production manager of a small company that manufactures a certain electronic component believes that playing some contemporary music in the production area will help reduce the number of nonconforming items produced. A group of workers with similar background (training, experience, etc.) are selected and five of them are assigned, at random, to work in the area while music is played. Then from the remaining group, four workers are randomly assigned to work in the usual way without music. The numbers of nonconforming items produced by the workers over a particular period of time are given below. Test to see if the median number of nonconforming items produced while music is played is less than that when no music is played.

Sample 1: Without music	*Sample 2: With music*
3, 4, 9, 10	1, 2, 5, 7, 8

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Solution Let samples 1 and 2 denote the X and Y sample, respectively. Assume the shift model and suppose that the null hypothesis to be tested is $M_X = M_Y$ against the alternative $M_Y < M_X$. Then, the P value for the median test is in the left tail. Since $N = 9$ is odd, $t = (9-1)/2 = 4$. The combined sample median is equal to 5 and thus $U = 2$. Using (4.4), the exact P value for the median test is

$$P(U \leqslant 2 \mid H_0) = \frac{\binom{4}{0}\binom{5}{4} + \binom{4}{1}\binom{5}{3} + \binom{4}{2}\binom{5}{2}}{\binom{9}{4}}$$

$$= 105/126 = 0.8333$$

There is not enough evidence in favor of the alternative H_1 and we do not reject H_0. The reader can verify using (4.7) that the normal approximation to the P value is $\Phi(0.975) = 0.8352$, leading to the same conclusion.

The MINITAB solution to Example 4.1 for the median test is shown below. For each group the MINITAB output gives the median and the interquartile range. Note that MINITAB does not calculate the P value for the exact median test but provides the chi-square approximation with $\mathrm{df} = 1$. The chi-square test statistic is the square of the Z statistic in (4.6), based on the normal approximation without a continuity correction. Calculations yield $Z^2 = 0.09$ and from Table B of the Appendix, the critical value at $\alpha = 0.05$ is found to be 3.84. Thus, the approximate test also fails to reject the null hypothesis. Using the chi-square approximation for these small sample sizes might not be advisable however. Using MINITAB, the right-tail probability corresponding to the observed value of 0.09 under the chi-square distribution with $df = 1$ is 0.7642 and this is in fact the P value shown in the printout. MINITAB also provides a 95% confidence interval for each group median (based on the sign test and interpolation) and for the difference in the medians. The two individual confidence intervals overlap, and the interval for the difference of medians includes zero, suggesting that the corresponding population medians are the same. The MINITAB output does not show the result of a median test but it does show a confidence interval for the difference between the medians based on a median test; this will be discussed next. For these data, the confidence interval is (−4.26, 8.26) which includes zero and this suggests, as before, that the population medians are not different.

```
            ********************************
            MINITAB SOLUTION TO EXAMPLE 4.1
            ********************************

Mood Median Test: C2 versus C1

Mood median test for C2

Chi-Square = 0.09    DF = 1    P = 0.764

                                          Individual 95.0% CIs
C1   N<=   N>   Median    Q3-Q1   ------+---------+---------+---------+
1    2     2     6.50     6.50           (-------------+-------------)
2    3     2     5.00     6.00    (---------------+-----------)
                                   ------+---------+---------+---------+
                                        2.5       5.0       7.5      10.0
Overall median = 5.00
* NOTE * Levels with < 6 observations have confidence < 95.0%

A 95.0% CI for median(1) - median(2): (-4.26,8.26)
```

The STATXACT solution is shown next. STATXACT at present does not provide the exact P value for the two-sample case directly. However, a little programming can be used to find the exact P value as shown below. Also, note that STATXACT bases the test on the chi-square statistic and not on the count U.

```
              ********************************
              STATXACT SOLUTION TO EXAMPLE 4.1
              ********************************

MEDIAN TEST

Statistics based on the observed one-way layout:

        Number of groups       = 2
        Number of observations = 9
        The overall median     = 5.000
        Observed Statistic     = 0.09000

Asymptotic p-value: (based on Chi-Square distribution with 1 df )
     Pr { CH(X) .GE. 0.09000 } =  0.7642
```

Finally, the SAS output is shown. SAS determines S, the number of observations that are above the combined sample median, for the sample with the smaller sample size. In our case the X sample has the smaller sample size and $S = 2$. According to SAS documentation, a one-sided P value is calculated as $P_1 = P(\text{Test Statistic} \geqslant S \mid H_0)$ if

$S >$ Mean, whereas if $S \leqslant$ Mean, the one-sided P value is calculated as $P_1 = P(\text{Test Statistic} \leqslant S \mid H_0)$. The mean of the median test statistic under H_0 is mt/N which equals $4(4)/9 = 1.78$ for our example. Thus, SAS calculates the exact P value in the upper tail as $P(S \geqslant 2 \mid H_0) = 81/126 = 0.6429$. This equals $1 - P(U \leqslant 1 \mid H_0)$ and thus does not agree with our hand calculations. However, on the basis of S we reach the same conclusion of not rejecting H_0, made earlier on the basis of U. For the normal approximation to the P value, PROC NPAR1WAY calculates the Z statistic by the formula $Z = (S - mt/N)/\sqrt{mnt(N-t)/N^2(N-1)}$ and incorporates a continuity correction unless one specifies otherwise. As with the exact P value, the SAS P value under the normal approximation also does not agree with our hand calculation based on U.

```
***************************************
SAS PROGRAM AND SOLUTION TO EXAMPLE 4.1
***************************************
```

Program:

```
data example;
      input sample number @@;
      datalines;
   1 3   1 4   1 9   1 10   2 1   2 2   2 5   2 7   2  8
   proc npar1way median data=example;
      class sample;
      var number;
      exact;
   run;
```

Output:

```
                       The NPAR1WAY Procedure

  Median Scores (Number of Points Above Median) for Variable number
                    Classified by Variable sample
```

sample	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	4	2.0	1.777778	0.785674	0.50
2	5	2.0	2.222222	0.785674	0.40

```
                    Median Two-Sample Test

         Statistic (S)              2.0000
```

```
Normal Approximation
Z                              0.2828
One-Sided Pr >  Z              0.3886
Two-Sided Pr > |Z|             0.7773

Exact Test
One-Sided Pr >=  S             0.6429
Two-Sided Pr >= |S - Mean|     1.0000

        Median One-Way Analysis

    Chi-Square          0.0800
    DF                       1
    Pr > Chi-Square     0.7773
```

CONFIDENCE-INTERVAL PROCEDURE

Assuming the shift model, the median-test procedure can be adapted to yield a confidence-interval estimate for the shift parameter in the following manner. Suppose that the two populations are identical in every way except for their medians. Denote these unknown parameters by M_X and M_Y, respectively, and the difference $M_Y - M_X$ by θ. In the shift model $F_Y(x) = F_X(x - \theta)$, the parameter θ represents the difference $F_Y^{-1}(p) - F_X^{-1}(p)$ between any two quantiles (of order p) of the two populations. In the present case, however, we assume that $\theta = M_Y - M_X$, the difference between the two medians ($p = 0.5$). From the original samples, if θ were known we could form the derived random variables

$$X_1, X_2, \ldots, X_m \qquad \text{and} \qquad Y_1 - \theta, Y_2 - \theta, \ldots, Y_n - \theta$$

and these would constitute samples from identical populations or, equivalently, a single sample of size $N = m + n$ from the common population. Then according to the median-test criterion with significance level α, the null hypothesis of identical distributions would be accepted for these derived samples if U, the number of X observations less than the median of the combined sample of derived observations, lies in the interval $c + 1 \leqslant U \leqslant c' - 1$. Recall that the rejection region against the two-sided alternative $\theta \neq 0$ is $U \leqslant c$ or $U \geqslant c'$. The integers c and c' are chosen such that

$$\sum_{u=0}^{c} \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}} + \sum_{u=c'}^{t} \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}} \leqslant \alpha \tag{4.8}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where $t = N/2$ or $(N-1)/2$ according as N is even or odd. Since H_0 is accepted for all U values in the interval $c+1 \leqslant U \leqslant c'-1$, and this acceptance region has probability $1-\alpha$ under H_0, a $100(1-\alpha)\%$ confidence-interval estimate for θ consists of all values of θ for which the derived sample observations yield values of U which lie in the acceptance region. This process of obtaining a confidence interval for a parameter from the acceptance region (of a test of hypothesis) is called inverting the acceptance region (of the test), and the confidence interval thus obtained is referred to as a *test-based confidence interval*.

To explicitly find the confidence interval, that is, the range of θ corresponding to the acceptance region $c+1 \leqslant U \leqslant c'-1$, we first order the two derived samples separately from smallest to largest as

$$X_{(1)}, X_{(2)}, \ldots, X_{(m)} \qquad \text{and} \qquad Y_{(1)}-\theta, Y_{(2)}-\theta, \ldots, Y_{(n)}-\theta$$

The t smallest observations of the $N = m+n$ total number are made up of exactly i X and $t-i$ Y variables if each observation of the set

$$X_{(1)}, \ldots, X_{(i)}, Y_{(1)}-\theta, \ldots, Y_{(t-i)}-\theta$$

is less than each observation of the set

$$X_{(i+1)}, \ldots, X_{(m)}, Y_{(t-i+1)}-\theta, \ldots, Y_{(n)}-\theta$$

The value of i is at least $c+1$ if and only if for $i = c+1$, the largest X in the first set is less than the smallest Y in the second set, that is, $X_{(c+1)} < Y_{(t-c)} - \theta$. Arguing similarly, $X_{(c')} > Y_{(t-c'+1)} - \theta$ can be seen to be a necessary and sufficient condition for having at most $c'-1$ X observations among the t smallest of the total N (in this case the largest Y in the first set must be smaller than the smallest X in the second set). Therefore, the acceptance region for the median test corresponding to the null hypothesis of no difference between the two distributions (with respect to location) at significance level α can be equivalently written as

$$X_{(c+1)} < Y_{(t-c)} - \theta \qquad \text{and} \qquad X_{(c')} > Y_{(t-c'+1)} - \theta$$

or as

$$Y_{(t-c)} - X_{(c+1)} > \theta \qquad \text{and} \qquad Y_{(t-c'+1)} - X_{(c')} < \theta$$

The desired confidence interval $(Y_{(t-c'+1)} - X_{(c')}, Y_{(t-c)} - X_{(c+1)})$ follows from the last two inequalities. Now, using (4.8),

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$\begin{aligned} 1 - \alpha &= P(c + 1 \leqslant U \leqslant c' - 1 \mid H_0) \\ &= P(c + 1 \leqslant U \leqslant c' - 1 \mid \theta = 0) \\ &= P(Y_{(t-c'+1)} - X_{(c')} < \theta < Y_{(t-c)} - X_{(c+1)} | \theta = 0) \end{aligned}$$

Since the last equality is also true for all values of θ, we can make the statement

$$P(Y_{(t-c'+1)} - X_{(c')} < \theta < Y_{(t-c)} - X_{(c+1)}) = 1 - \alpha$$

where c and c' are found from (4.8). Thus the endpoints of the confidence interval estimate for θ corresponding to Mood's median test are found simply from some order statistics of the respective random samples.

Example 4.2 We calculate the 95% confidence interval for the median difference for the data in Example 4.1. In order to find the constants c and c', we need to calculate the null distribution of U, using (4.3) for $m = 4, n = 5, t = 4$. The results are shown in Table 4.1. If we take $c = 0$ and $c' = 4$, then (4.8) equals 0.04762 so that the confidence interval for $\theta = M_Y - M_X$ is $(Y_{(1)} - X_{(4)}, Y_{(4)} - X_{(1)})$ with exact level 0.95238. Numerically, the intervals is $(-9,4)$. Also, the 95.238% confidence interval for $\theta = M_X - M_Y$ is $(-4,9)$. Note that the MINITAB output given before states "A 95.0% CI for median(1) – median(2): $(-4.26, 8.26)$." This is based on the median test but c and c' are calculated using the normal approximation. The results are quite close.

It may be noted that the median test is a member of a more general class of nonparametric two-sample tests, called *precedence tests*. Chakraborti and van der Laan (1996) provided an overview of the literature on these tests. A precedence test is based on a statistic W_r which denotes the number of Y observations that precede the rth-order statistic from the X sample $X_{(r)}$ (alternatively, one can use the number of X's that precede, say, $Y_{(s)}$). It can be seen, for example, that $W_r < w$ if and only if $X_{(r)} < Y_{(w)}$ so that a precedence test based on W_r can be interpreted in terms of two order statistics, one from each sample. The test is implemented by first choosing r, and then determining w such that the size of the test is α. It can be shown that

Table 4.1 Null distribution of U for $m = 4, n = 5, t = 4$

u	0	1	2	3	4
$P(U = u)$	0.039683	0.31746	0.47619	0.15873	0.007937
	5/126	40/126	60/126	20/126	1/126

the null distribution of W_r is distribution free (see, for example, Problem 2.28), so that the precedence test is a distribution-free test. The median test a special case of a precedence test since as seen in the arguments for the confidence interval procedure (see also, for example, Pratt, 1964), we have $U \leqslant u - 1$ if and only if $X_{(u)} < Y_{(t-u+1)}$. Several precedence tests have been proposed in the literature and we will refer to some of some of them in this chapter.

POWER OF THE MEDIAN TEST

The power of the median test can be obtained as a special case of the power of a precedence test and we sketch the arguments for the more general case. The power of a precedence test can be obtained from the distribution of the precedence statistic W_r under the alternative hypothesis. It can be shown that for $i = 0, 1, \ldots, n$

$$P(W_r = i) = \frac{\binom{n}{i}}{B(r, m-r+1)} \times \int_0^1 \left[F_Y\left(F_X^{-1}(u)\right)\right]^i \left[1 - F_Y\left(F_X^{-1}(u)\right)\right]^{n-i} u^{r-1}(1-u)^{m-r}\, du \tag{4.9}$$

Thus, for a one-sided precedence test with rejection region $W_r < w_\alpha$, the power of the test is

$$\mathrm{Pw}(F_X, F_Y, m, n, \alpha) = \sum_{i=0}^{w_\alpha - 1} P(W_r = i) \tag{4.10}$$

where w_α is obtained from the size of the test; in other words, w_α is the largest integer such that

$$\sum_{s=0}^{w_\alpha - 1} P(W_r = s | H_0) \leqslant \alpha$$

To obtain the power of the median test, we just need to substitute suitable values in (4.10).

As an alternative development of power, note that under the assumption of the location model $F_Y(x) = F_X(x - \theta)$ the null hypothesis is that θ equals zero. From the confidence interval procedure developed above, we know that for the median test a necessary and sufficient condition for accepting this hypothesis is that the number zero be included in the random interval

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$[(Y_{(t-c'+1)} - X_{(c')}), (Y_{(t-c)} - X_{(c+1)})]$$

Thus, the power function of the median test in the location case is the probability that this interval does not cover zero when $\theta \neq 0$, that is,

$$\text{Pw}(\theta) = P(Y_{(t-c'+1)} - X_{(c')} > 0 \text{ or } Y_{(t-c)} - X_{(c+1)} < 0 \text{ when } \theta \neq 0)$$

These two events, call them A and B, are mutually exclusive as we now show. For any $c' > c$, it is always true that $X_{(c')} \geqslant X_{(c+1)}$ and $Y_{(t-c'+1)} = Y_{(t-[c'-1])} \leqslant Y_{(t-c)}$. Thus if A occurs, that is, $Y_{(t-c'+1)} > X_{(c')}$, we must also have $Y_{(t-c)} \geqslant Y_{(t-c'+1)} > X_{(c')} \geqslant X_{(c+1)}$ which makes $Y_{(t-c)} > X_{(c+1)}$, a contradiction in B. As a result, the power function can be expressed as the sum of two probabilities involving order statistics:

$$\text{Pw}(\theta) = P(Y_{(t-c'+1)} > X_{(c')}) + P(Y_{(t-c)} < X_{(c+1)})$$

Since the random variables X and Y are independent, the joint distribution of, say, $X_{(r)}$ and $Y_{(s)}$ is the product of their marginal distributions, which can be easily found using the methods of Chapter 2 for completely specified populations F_X and F_Y or, equivalently, F_X and θ since $F_Y(x) = F_X(x - \theta)$. In order to calculate the power function then, we need only evaluate two double integrals of the following type:

$$P(Y_{(s)} < X_{(r)}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_{(s)}}(y) f_{X_{(r)}}(x)\, dy\, dx$$

The power function for a one-sided test is simply one integral of this type. For large sample sizes, since the marginal distribution of any order statistic approaches the normal distribution and the order statistics $X_{(r)}$ and $Y_{(s)}$ are independent here, the distribution of their difference $Y_{(s)} - X_{(r)}$ approaches the normal distribution with mean and variance

$$E(Y_{(s)}) - E(X_{(r)}) \qquad \text{and} \qquad \text{var}(Y_{(s)}) + \text{var}(X_{(r)})$$

Given the specified distribution function and the results in Chapter 2, we can approximate these quantities by

$$E(X_{(r)}) = F_X^{-1}\left(\frac{r}{m+1}\right) \qquad E(Y_{(s)}) = F_X^{-1}\left(\frac{s}{n+1}\right)$$

$$\text{var}(X_{(r)}) = \frac{r(m-r+1)}{(m+1)^2(m+2)} \left\{ f_X\left[F_X^{-1}\left(\frac{r}{m+1}\right)\right]\right\}^{-2}$$

$$\text{var}(Y_{(s)}) = \frac{s(n-s+1)}{(n+1)^2(n+2)} \left\{ f_Y\left[F_X^{-1}\left(\frac{s}{n+1}\right)\right]\right\}^{-2}$$

and an approximation to the power function can be found using normal probability tables.

It is clear that computing the exact or even the asymptotic power of the median test is computationally involved. An easier alternative approach might be to use computer simulations, as was outlined for the sign and the signed rank test in Chapter 5. We leave the details to the reader.

The asymptotic efficiency of the median test relative to Student's t test for normal populations is $2/\pi = 0.637$ (see Chapter 13). As a test for location, this is relatively poor performance. The Mann-Whitney test, discussed in Section 6.6, has greater efficiency for normal populations.

6.5 THE CONTROL MEDIAN TEST

The median test, based on the number of X observations that precede the median of the combined samples, is a special case of a precedence test. A simple yet interesting alternative test is a *second precedence test*, based on the number of X (or Y) observations that precede the median of the Y (or X) sample. This is known as the *control median test* and is generally attributed to Mathisen (1943). The properties and various refinements of the test have been studied by Gart (1963), Gastwirth (1968), and Hettmansperger (1973), among others.

Without any loss of generality, suppose the Y sample is the control sample. The control median test is based on V, the number of X observations that precede the median of the Y observations. For simplicity let $n = 2r + 1$, so that the $(r + 1)$th-order statistic $Y_{(r+1)}$ is the median of the Y sample. Now $Y_{(r+1)}$ defines two nonoverlapping blocks $[-\infty, Y_{(r+1)}]$ and $(Y_{(r+1)}, \infty)$ in the sample, and the control median test is based on V, the number of X observations in the first block. It may be noted that V is equal to $mS_m(Y_{(r+1)}) = P_{(r+1)}$, called the *placement* of $Y_{(r+1)}$, the median of the Y sample, among the X observations.

As with the median test, the control median test can be used to test the null hypothesis H_0: $F_Y(x) = F_X(x)$ for all x against the general one-sided alternative that for example, the Y's are stochastically larger than the X's. In this case, the number of X's preceding $Y_{(r+1)}$ should be large and thus large values of V provide evidence against the null hypothesis in favor of the alternative. If we assume the shift model $F_Y(x) = F_X(x - \theta)$ then the problem reduces to testing the null hypothesis H_0: $\theta = 0$ against the one-sided alternative hypothesis H_1: $\theta > 0$ and the appropriate rejection region consists of large values of V.

In a similar manner it can be seen that for the one-sided alternative that the Y's are stochastically smaller than the X's or under the shift model H_1: $\theta < 0$, the rejection region should consist of small values of V.

Since the control median test is a precedence test, Problem 2.28(c) (with $n = 2r + 1$ and $i = r + 1$) gives the null distribution of the test statistic V as:

$$P[V = j \,|\, H_0] = \frac{\binom{m+r-j}{m-j}\binom{j+r}{j}}{\binom{m+2r+1}{m}} \tag{5.1}$$

for $j = 0, 1, \ldots, m$. This can be easily evaluated to find the exact P value corresponding to an observed value V_O or to calculate a critical value for a given level of significance α. Further, from Problem 2.28 (d), the null mean and the variance of V are

$$E(V) = \frac{m(r+1)}{(n+1)} = \frac{m}{2}$$

and

$$\begin{aligned}\operatorname{var}(V) &= \frac{m(m+n+1)(r+1)(n-r)}{(n+1)^2(n+2)}\\ &= \frac{m(m+2r+2)}{4(2r+3)} = \frac{m(m+n+1)}{4(n+2)}\end{aligned}$$

In general, V/m is an estimator of $F_X(M_Y) = q$, say, and in fact is a consistent estimator. Now, q is equal to 0.5 under H_0, is less than 0.5 if and only if $M_Y < M_X$, and is greater than 0.5 if and only if $M_Y > M_X$. Hence, like the median test, a test based on V is especially sensitive to differences in the medians.

Gastwirth (1968) showed that when m, n tend to infinity such that m/n tends to ϕ, and some regularity conditions are satisfied, $\sqrt{m}(V/m - q)$ is asymptotically normally distributed with mean zero and variance

$$q(1-q) + \phi\frac{f_X^2(M_Y)}{4f_Y^2(M_Y)}$$

Under H_0 we have $q = 0.5, f_X = f_Y$, and $M_X = M_Y$, so that the asymptotic null mean and variance of V are $m/2$ and $m(m+n)/4n$, respectively. Thus, under H_0, the asymptotic distribution of

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$Z = \frac{V - m/2}{\sqrt{m(m+n)/4n}} = \frac{\sqrt{n}(2V - m)}{\sqrt{m(m+n)}}$$

is approximately standard normal.

Suppose we are interested in testing only the equality of the two medians (or some other quantiles) and not the entire distributions. In the context of location-scale models, the null hypothesis may concern only the equality of location parameters, without assuming that the scale parameters are equal. By analogy with a similar problem in the context of normal distributions, this is a nonparametric Behrens-Fisher problem. Note that under the current null hypothesis we still have $q = 0.5$ but the ratio $f_X(M)/f_Y(M)$, where M is the common value of the medians under the null hypothesis, does not necessarily equal one. This implies that in order to use the control median test for this problem we need to estimate this ratio of the two densities at M. Once a suitable estimate is found, the asymptotic normality of V can be used to construct an approximate test of significance. Several authors have studied this problem, including Pothoff (1963), Hettmansperger (1973), Hettmansperger and Malin (1975), Schlittgen (1979), Smit (1980), and Fligner and Rust (1982).

CURTAILED SAMPLING

The control median test, or more generally any precedence test, is particularly useful in life-testing experiments, where observations are naturally time ordered and collecting data is expensive. In such experiments, the precedence tests allow a decision to be made about the null hypothesis as soon as a preselected ordered observation becomes available. Thus the experiment can be terminated (or the sampling procedure can be curtailed) before all of the data have been collected, and precedence tests have the potential of saving time and resources. Note that the decision made on the basis of the "curtailed sample" is always the same as it would have been if all observations had been available.

As an illustration consider testing $H_0\colon q = 0.5$ again the one-sided alternative $H_1\colon q < 0.5$. Using the normal approximation, the control median test would reject H_0 in favor of H_1 if $V \leqslant d$, where

$$d = \frac{m}{2} + z_{\alpha/2}\left[\frac{m(m+n)}{4n}\right]^{1/2} \tag{5.2}$$

or equivalently if the median of the Y sample of size $2r + 1$ satisfies

$$Y_{(r+1)} \leqslant X_{(d)}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where d is the solution from (5.2) after rounding down.

This restatement of the rejection region in terms of the X- and Y-order statistics clearly shows that a decision can be reached based on the control median test as soon as the median of the Y sample or the dth order statistic of the X sample is observed, whichever comes first. The index d is fixed by the size of the test and can be obtained exactly from the null distribution of V or from the normal approximation shown in (5.2). The null hypothesis is rejected if the median of the Y sample is observed before the dth-order statistic of the X sample; otherwise the null hypothesis is accepted.

Gastwirth (1968) showed that in large samples the control median test always provides an earlier decision than the median test for both the one- and two-sided alternatives. For related discussions and other applications, see Pratt and Gibbons (1981) and Young (1973).

POWER OF THE CONTROL MEDIAN TEST

Since the median test and the control median test are both precedence tests, the power function of the control median test can be obtained in a manner similar to that for the median test. If the alternative is one-sided, say $q < 0.5$, the control median test rejects the null hypothesis $q = 0.5$ at significance level α if $V \leqslant d_\alpha$. Hence the power of this control median test is $P(V \leqslant d_\alpha \mid q < 0.5)$. However, the event $V \leqslant d_\alpha$ is equivalent to $Y_{(r+1)} \leqslant X_{(d^*)}$, where $d^* = [d_\alpha / m]$, and therefore the power of the test is simply

$$\text{Pw}(\theta) = P[Y_{(r+1)} \leqslant X_{(d^*)} | q < 0.5]$$

where $q = F_X(M_Y)$ and d_α satisfies $P(V \leqslant d_\alpha | H_0) \leqslant \alpha$. Note that the power of the control median test depends on this composite function $q = F_X(M_Y)$, in addition to α, m and n, and q is not necessarily equal to the difference of medians, not even under the shift model. The quantity q can be thought of as a parameter that captures possible differences between the Y and the X distributions.

DISCUSSION

The control median test provides an alternative to the median test. The ARE relative to the median test is one regardless of the continuous parent distributions, and in this sense the test is as efficient as the median test in large samples. In fact, the efficacy of the control median test is a symmetric function of the two sample sizes, which implies that "the designation of the numbering of the samples is immaterial as far as asymptotic efficiency is concerned" (Gart, 1963).

Because the ARE between the control median test and the median test is equal to one, neither one is preferable based on this criterion. However, if some other measure of efficiency is used, interesting distinctions can be made between the two tests. For these and other related results see, for example, Killeen, Hettmansperger, and Sievers (1972) and the references therein. As noted earlier, the control median test and the median tests are special cases of precedence tests. An overview of precedence tests for various problems can be found in Chakraborti and van der Laan (1996, 1997).

The control median test is based on V, the number of X values that precede the median of the Y's. Writing $q = F_X(M_Y)$, the appropriate rejection regions for a nominal significance level α are shown in the following table along with expressions for P values where d_α and d'_α are, respectively, the largest and smallest integers such that $P(V \leqslant d_\alpha|H_0) \leqslant \alpha, P(V \geqslant d'_\alpha|H_0) \leqslant \alpha$ and d and d' $(d < d')$ are two positive integers such that

$$P(V \leqslant d \,|\, H_0) + P(V \geqslant d' \,|\, H_0) \leqslant \alpha$$

The exact critical values or P values can be easily found directly using (5.1) or from the tables of the hypergeometric distribution (see Problem 6.10). In view of the simple form of the null distribution, it might be easier to calculate a P value corresponding to an observed value of V. Note that unlike the median test, there is no difficulty here in assigning probability in the tails with two-tailed tests since the distribution of V is symmetric under H_0, that is, $P(V = j \,|\, H_0) = P(V = m - j \,|\, H_0)$ for all j.

Alternative	*Rejection region*	*P value*
$Y \overset{ST}{>} X$, $q > 0.5$ $\theta > 0$ or $M_Y > M_X$	$V \leqslant d'_\alpha$	$P(V \geqslant V_O|H_0) = \sum_{j=V_0}^{m} P[V = j|H_0]$
$Y \overset{ST}{<} X$, $q < 0.5$ $\theta < 0$ or $M_Y < M_X$	$V \leqslant d_\alpha$	$P(V \leqslant V_O|H_0) = \sum_{j=0} V_0 P[V = j|H_0]$
$\theta \neq 0$ or $M_Y \neq M_X$ $q \neq 0.5$	$V \leqslant d$ or $V \geqslant d'$	2 (smaller of above)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

In practice, the asymptotic distribution is useful for finding the critical values or approximating the P value. For example, if the alternative is $q \leqslant 0.5$, an approximation to the P value of the test with a continuity correction is

$$\Phi\left[\frac{\sqrt{n}(2v - m + 1)}{\sqrt{m(m+n)}}\right] \tag{5.3}$$

APPLICATIONS

Example 5.1 We illustrate the control median test using the data in Example 4.1.

Solution Again, let samples 1 and 2 denote the X and the Y sample, respectively and assume that the null hypothesis to be tested is $M_X = M_Y$ $(q = 0.5)$ against the alternative $M_Y < M_X$ $(q < 0.5)$ under the shift model. Thus, the P value for the control median test is also in the left tail. The median of the Y sample is 5 and thus $V = 2$. Using (5.1) with $m = 4, n = 5$, and $r = 2$, the exact P value for the control median test is

$$\begin{aligned} P(V \leqslant 2 \,|\, H_0) &= \frac{\binom{6}{4}\binom{2}{0} + \binom{5}{3}\binom{3}{1} + \binom{4}{2}\binom{4}{2}}{\binom{9}{4}} \\ &= 81/126 = 0.6429 \end{aligned}$$

Hence there is not enough evidence against the null hypothesis in favor of the alternative. Also, from (5.3) the normal approximation to the P value is found to be $\Phi(0.37) = 0.6443$ and the approximate test leads to the same decision as the exact test.

For the median test, the combined sample median is equal to the Y sample median and $U = 2$. Now using (4.4) the exact P value for the median test is

$$\begin{aligned} P(U \leqslant 2 \,|\, H_0) &= \frac{\binom{4}{0}\binom{5}{4} + \binom{4}{1}\binom{5}{3} + \binom{4}{2}\binom{5}{2}}{\binom{9}{4}} \\ &= 105/126 = 0.8333 \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and once again not enough evidence is available in favor of the alternative H_1. Using (4.7), the normal approximation to the P value is obtained as $\Phi(0.975) = 0.8352$, leading to the same conclusion.

6.6 THE MANN-WHITNEY *U* TEST

Like the Wald-Wolfowitz runs test in Section 6.2, the Mann-Whitney U test (Mann and Whitney, 1947) is based on the idea that the particular pattern exhibited when the X and Y random variables are arranged together in increasing order of magnitude provides information about the relationship between their populations. However, instead of measuring the tendency to cluster by the total number of runs, the Mann-Whitney criterion is based on the magnitudes of, say , the Y's in relation to the X's, that is, the position of the Y's in the combined ordered sequence. A sample pattern of arrangement where most of the Y's are greater than most of the X's, or vice versa, or both, would be evidence against a random mixing and thus tend to discredit the null hypothesis of identical distributions.

The *Mann-Whitney U test statistic* is defined as the number of times a Y precedes an X in the combined ordered arrangement of the two independent random samples

$$X_1, X_2, \ldots, X_m \qquad \text{and} \qquad Y_1, Y_2, \ldots, Y_n$$

into a single sequence of $m + n = N$ variables increasing in magnitude. We assume that the two samples are drawn from continuous distributions, so that the possibility $X_i = Y_j$ for some i and j need not be considered. If the mn indicator random variables are defined as

$$D_{ij} = \begin{cases} 1 & \text{if } Y_j < X_i \\ & \quad \text{for } i = 1, 2, \ldots, m, \ j = 1, 2, \ldots, n \\ 0 & \text{if } Y_j > X_i \end{cases} \tag{6.1}$$

a symbolic representation of the Mann-Whitney U statistic is

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} \tag{6.2}$$

The logical rejection region for the one-sided alternative that the Y's are stochastically larger than the X's,

$$H_1\colon F_Y(x) \leqslant F_X(x) \qquad \text{strict inequality for some } x$$

would clearly be small values of U. The fact that this is a consistent test criterion can be shown by investigating the convergence of U/mn

to a certain parameter where H_0 can be written as a statement concerning the value of that parameter.

For this purpose, we define

$$p = P(Y < X) = \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_Y(y) f_X(x)\, dy\, dx = \int_{-\infty}^{\infty} F_Y(x)\, dF_X(x) \tag{6.3}$$

and the hypothesis testing problem can be redefined in terms of the parameter p. If H_0: $F_Y(x) = F_X(x)$ for all x is true, then

$$p = \int_{-\infty}^{\infty} F_X(x)\, dF_X(x) = 0.5 \tag{6.4}$$

If, for example, the alternative hypothesis is H_1: $F_Y(x) \leqslant F_X(x)$ that is $Y \overset{\text{ST}}{>} X$, then H_1: $p \leqslant 0.5$ for all x and $p < 0.5$ for some x. Thus the null hypothesis of identical distributions can be parameterized to H_0: $p = 0.5$ and the alternative hypothesis to H_1: $p < 0.5$.

The mn random variables defined in (6.1) are Bernoulli variables, with moments

$$E(D_{ij}) = E(D_{ij}^2) = p \qquad \operatorname{var}(D_{ij}) = p(1-p) \tag{6.5}$$

For the joint moments we note that these random variables are not independent whenever the X subscripts or the Y subscripts are common, so that

$$\begin{aligned} &\operatorname{cov}(D_{ij}, D_{hk}) = 0 \qquad \text{for } i \neq h \text{ and } j \neq k \\ &\underset{j \neq k}{\operatorname{cov}}(D_{ij}, D_{ik}) = p_1 - p^2 \qquad \underset{i \neq h}{\operatorname{cov}}(D_{ij}, D_{hj}) = p_2 - p^2 \end{aligned} \tag{6.6}$$

where the additional parameters introduced are

$$\begin{aligned} p_1 &= P(Y_j < X_i \cap Y_k < X_i) \\ &= P(Y_j \text{ and } Y_k < X_i) \\ &= \int_{-\infty}^{\infty} [F_Y(x)]^2\, dF_X(x) \end{aligned} \tag{6.7}$$

and

$$\begin{aligned} p_2 &= P(X_i > Y_j \cap X_h > Y_j) \\ &= P(X_i \text{ and } X_h > Y_j) \\ &= \int_{-\infty}^{\infty} [1 - F_X(y)]^2\, dF_Y(y) \end{aligned} \tag{6.8}$$

Since U is defined in (6.2) as a linear combination of these mn random variables, the mean and variance of U are

$$E(U) = \sum_{i=1}^{m} \sum_{j=1}^{n} E(D_{ij}) = mnp \tag{6.9}$$

$$\begin{aligned}\operatorname{var}(U) &= \sum_{i=1}^{m} \sum_{j=1}^{n} \operatorname{var}(D_{ij}) + \sum_{i=1}^{m} \sum\sum_{1 \leqslant j \neq k \leqslant n} \operatorname{cov}(D_{ij}, D_{ik}) \\ &\quad + \sum_{j=1}^{n} \sum\sum_{1 \leqslant i \neq h \leqslant m} \operatorname{cov}(D_{ij}, D_{hj}) \\ &\quad + \sum\sum_{1 \leqslant i \neq h \leqslant m} \sum\sum_{1 \leqslant j \neq k \leqslant n} \operatorname{cov}(D_{ij}, D_{hk})\end{aligned} \tag{6.10}$$

Now substituting (6.5) and (6.6) in (6.10), this variance is

$$\begin{aligned}\operatorname{var}(U) &= mnp(1-p) + mn(n-1)(p_1 - p^2) + nm(m-1)(p_2 - p^2) \\ &= mn[p - p^2(N-1) + (n-1)p_1 + (m-1)p_2]\end{aligned} \tag{6.11}$$

Since $E(U/mn) = p$ and $\operatorname{var}(U/mn) \to 0$ as $m, n \to \infty$, U/mn is a consistent estimator of p. Based on the method described in Chapter 1, the Mann-Whitney test can be shown to be consistent in the following cases:

Alternative		*Rejection region*
$p < 0.5$	$F_Y(x) \leqslant F_X(x)$	$U - mn/2 < k_1$
$p > 0.5$	$F_Y(x) \geqslant F_X(x)$	$U - mn/2 > k_2$
$p \neq 0.5$	$F_Y(x) \neq F_X(x)$	$U - mn/2 > k_3$

(6.12)

In order to determine the size α critical regions of the Mann-Whitney test, we must now find the null probability distribution of U. Under H_0, each of the $\binom{m+n}{m}$ arrangements of the random variables into a combined sequence occurs with equal probability, so that

$$f_U(u) = P(U = u) = \frac{r_{m,n}(u)}{\dbinom{m+n}{m}} \tag{6.13}$$

where $r_{m,n}(u)$ is the number of distinguishable arrangements of the m X and n Y random variables such that in each sequence the number of times a Y precedes an X is exactly u. The values of u for which $f_U(u)$ is nonzero range between zero and mn, for the two most extreme orderings in which every x precedes every y and every y precedes every x, respectively. We first note that the probability distribution of U is symmetric about the mean $mn/2$ under the null hypothesis. This property may be argued as follows. For every particular arrangement z of the m x and n y letters, define the conjugate arrangement z' as the sequence z written backward. In other words, if z denotes a set of numbers written from smallest to largest, z' denotes the same numbers written from largest to smallest. Every y that precedes an x in z then follows that x in z', so that if u is the value of the Mann-Whitney statistic for $z, mn - u$ is the value for z'. Therefore under H_0, we have $r_{m,n}(u) = r_{m,n}(mn - u)$ or, equivalently,

$$P\left(U - \frac{mn}{2} = u\right) = P\left(U = \frac{mn}{2} + u\right)$$
$$= P\left[U = mn - \left(\frac{mn}{2} + u\right)\right] = P\left(U - \frac{mn}{2} = -u\right)$$

Because of this symmetry property, only lower tail critical values need be found for either a one- or two-sided test. We define the random variable U' as the number of times an X precedes a Y or, in the notation of (6.1),

$$U' = \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - D_{ij})$$

and redefine the rejection regions for size α tests corresponding to (6.12) as follows:

	Alternative	*Rejection region*
$p < 0.5$	$F_Y(x) \leqslant F_X(x)$	$U \leqslant c_\alpha$
$p > 0.5$	$F_Y(x) \geqslant F_X(x)$	$U' \leqslant c_\alpha$
$p \neq 0.5$	$F_Y(x) \neq F_X(x)$	$U \leqslant c_{\alpha/2}$ or $U' \leqslant c_{\alpha/2}$

To determine the number c_α for any m and n, we can enumerate the cases starting with $u = 0$ and work up until at least $\alpha\binom{m+n}{m}$ cases are counted. For example, for $m = 4,\ n = 5$, the arrangements with the smallest values of u, that is, where most of the X's are smaller than most of the Y's, are shown in Table 6.1. The rejection regions for this one-sided test for nominal significance levels of 0.01 and 0.05 would then be $U \leqslant 0$ and $U \leqslant 2$, respectively.

Table 6.1 Generation of the left-tail *P* values of *U* for $m = 4,\ n = 5$

Ordering	u	
XXXXYYYYY	0	
XXXYXYYYY	1	$P(U \leqslant 0) = 1/126 = 0.008$
XXYXXYYYY	2	$P(U \leqslant 1) = 2/126 = 0.016$
XXXYYXYYY	2	$P(U \leqslant 2) = 4/126 = 0.032$
XYXXXYYYY	3	$P(U \leqslant 3) = 7/126 = 0.056$
XXYXYXYYY	3	
XXXYYYXYY	3	

Even though it is relatively easy to guess which orderings will lead to the smallest values of u, $\binom{m+n}{m}$ increases rapidly as m, n increase. Some more systematic method of generating critical values is needed to eliminate the possibility of overlooking some arrangements with u small and to increase the feasible range of sample sizes and significance levels for constructing tables. A particularly simple and useful recurrence relation can be derived for the Mann-Whitney statistic. Consider a sequence of $m + n$ letters being built up by adding a letter to the right of a sequence of $m + n - 1$ letters. If the $m + n - 1$ letters consist of m x and $n - 1$ y letters, the extra letter must be a y. But if a y is added to the right, the number of times a y precedes an x is unchanged. If the additional letter is an x, which would be the case for $m - 1$ x and n y letters in the original sequence, all of the y's will precede this new x and there are n of them, so that u is increased by n. These two possibilities are mutually exclusive. Using the notation of (6.13) again, this recurrence relation can be expressed as

$$r_{m,n}(u) = r_{m,n-1}(u) + r_{m-1,n}(u-n)$$

and

$$f_U(u) = p_{m,n}(u) = \frac{r_{m,n-1}(u) + r_{m-1,n}(u-n)}{\binom{m+n}{m}}$$
$$= \frac{n}{m+n}\frac{r_{m,n-1}(u)}{\binom{m+n-1}{n-1}} + \frac{m}{m+n}\frac{r_{m-1,n}(u-n)}{\binom{m+n-1}{m-1}}$$

or

$$(m+n)p_{m,n}(u) = np_{m,n-1}(u) + mp_{m-1,n}(u-n) \tag{6.14}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

This recursive relation holds for all $u = 0, 1, 2, \ldots, mn$ and all integer-valued m and n if the following initial and boundary conditions are defined for all $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

$$\begin{array}{ll} r_{ij}(u) = 0 & \text{for all } u < 0 \\ r_{i,0}(0) = 1 & r_{0,i}(0) = 1 \\ r_{i,0}(u) = 0 & \text{for all } u \neq 0 \\ r_{0,i}(u) = 0 & \text{for all } u \neq 0 \end{array}$$

If the sample with fewer observations is always labeled the X sample, tables are needed only for $m \leqslant n$ and left-tail critical points. Such tables are widely available, for example in Auble (1953) or Mann and Whitney (1947).

When m and n are too large for the existing tables, the asymptotic probability distribution can be used. Since U is the sum of identically distributed (though dependent) random variables, a generalization of the central-limit theorem allows us to conclude that the null distribution of the standardized U approaches the standard normal as $m, n \to \infty$ in such a way that m/n remains constant (Mann and Whitney, 1947). To make use of this approximation, the mean and variance of U under the null hypothesis must be determined. When $F_Y(x) = F_X(x)$, the integrals in (6.7) and (6.8) are evaluated as $p_1 = p_2 = 1/3$. Substituting these results in (6.9) and (6.11) along with the value $p = 1/2$ from (6.4) gives

$$E(U \mid H_0) = \frac{mn}{2} \qquad \text{var}(U \mid H_0) = \frac{mn(N+1)}{12} \tag{6.15}$$

The large-sample test statistic then is

$$Z = \frac{U - mn/2}{\sqrt{mn(N+1)/12}}$$

whose distribution is approximately standard normal. This approximation has been found reasonably accurate for equal sample sizes as small as 6. Since U can assume only integer values, a continuity correction of 0.5 may be used.

THE PROBLEM OF TIES

The definition of U in (6.2) was adopted for presentation here because most tables of critical values are designed for use in the way described above. Since D_{ij} is not defined for $X_i = Y_j$, this expression does not allow for the possibility of ties across samples. If ties occur within one or both of the samples, a unique value of U is obtained. However, if one or more X is tied with one or more Y, our definition requires that the ties be broken in some way. The conservative approach may be adopted, which means that all ties are broken in all possible ways and the largest resulting value of U (or U') is used in reaching the decision. When there are many ties (as might be the case when each random variable can assume only a few ordered values such as very strong, strong, weak, very weak), an alternative approach may be preferable.

A common definition of the Mann-Whitney statistic which does allow for ties (see Problem 5.1) is

$$U_T = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

where

$$D_{ij} = \begin{cases} 1 & \text{if } X_i > Y_j \\ 0.5 & \text{if } X_i = Y_j \\ 0 & \text{if } X_i < Y_j \end{cases} \tag{6.16}$$

However, it is often more convenient to work with

$$D_{ij}^* = \begin{cases} 1 & \text{if } X_i > Y_j \\ 0 & \text{if } X_i = Y_j \\ -1 & \text{if } X_i < Y_j \end{cases}$$

If the two parameters p^+ and p^- are defined as

$$p^+ = P(X > Y) \qquad \text{and} \qquad p^- = P(X < Y)$$

U_T may be considered as an estimate of its mean

$$E(U_T) = mn(p^+ - p^-)$$

A standardized U_T is asymptotically normally distributed. Since under the null hypothesis $p^+ = p^-$, we have $E(U_T \mid H_0) = 0$ whether ties occur or not. The presence of ties does affect the variance, however. The variance of U_T conditional upon the observed ties can be calculated in a manner similar to the steps leading to (6.11) by introducing some additional parameters. Then a correction for ties can be incorporated in the standardized variable used for the test statistic. The result is

$$\text{var}(U_T \mid H_0) = \frac{mn(N+1)}{12}\left[1 - \frac{\sum(t^3 - t)}{N(N^2-1)}\right] \tag{6.17}$$

where t denotes the multiplicity of a tie and the sum is extended over all sets of t ties. The details will be left to the reader as an exercise [or see Noether, 1967, pp. 32–35].

CONFIDENCE-INTERVAL PROCEDURE

If the populations from which the X and Y samples are drawn are identical in every respect except location, say $F_Y(x) = F_X(x - \theta)$ for all x and some θ, we say that the Y population is the same as the X population but shifted by an amount θ, which may be either positive or negative, the sign indicating the direction of the shift. We wish to use the Mann-Whitney test procedure to find a confidence interval for θ, the amount of shift. Under the assumption that $F_Y(x) = F_X(x - \theta)$ for all x and some θ, the sample observations $X_1, X_2, \ldots, X_m$ and $Y_1 - \theta$, $Y_2 - \theta, \ldots, Y_n - \theta$ come from identical populations. By a confidence interval for θ with confidence coefficient $1 - \alpha$ we mean the range of values of θ for which the null hypothesis of identical populations will be accepted at significance level α.

To apply the Mann-Whitney procedure to this problem, the random variable U now denotes the number of times a $Y - \theta$ precedes an X, that is, the number of pairs $(X_i, Y_j - \theta)$, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, for which $X_i > Y_j - \theta$, or equivalently, $Y_j - X_i < \theta$. If a table of critical values for a two-sided U test at level α gives a rejection region of $U \leqslant k$, say, we reject H_0 when no more than k differences $Y_j - X_i$ are less than the value θ, and accept H_0 when more than k differences are less than θ. The total number of differences $Y_j - X_i$ is mn. If these differences are ordered from smallest to largest according to actual (not absolute) magnitude, denoted by $D_{(1)}, D_{(2)}, \ldots, D_{(mn)}$, there are exactly k differences less than θ if θ is the $(k+1)$st-ordered difference, $D_{(k+1)}$. Any number exceeding this $(k+1)$st difference will produce more than k differences less than θ. Therefore, the lower limit of the confidence interval for θ is $D_{(k+1)}$. Similarly, since the probability distribution of U is symmetric, an upper confidence limit is given by that difference which is $(k+1)$st from the largest, that is, $D_{(mn-k)}$. The confidence interval with coefficient $1 - \alpha$ then is

$$D_{(k+1)} < \theta < D_{(mn-k)}$$

The procedure is illustrated by the following numerical example. Suppose that $m = 3, n = 5, \alpha = 0.10$. By simple enumeration, we find $P(U \leqslant 1) = 2/56 = 0.036$ and $P(U \leqslant 2) = 4/56 = 0.071$, and so the critical value when $\alpha/2 = 0.05$ is 1, with the exact probability of a type I error 0.072. The confidence interval will then be $D_{(2)} < \theta < D_{(14)}$.

Table 6.2 Confidence-interval calculations

$y_j - 1$	$y_j - 6$	$y_j - 7$
1	−4	−5
3	−2	−3
8	3	2
9	4	3
11	6	5

Suppose that the sample data are X: 1, 6, 7; Y: 2, 4, 9, 10, 12. In order to find $D_{(2)}$ and $D_{(14)}$ systematically, we first order the x and y data separately, then subtract from each y, starting with the smallest y, the successive values of x as shown in Table 6.2, and order the differences. The interval here is $-4 < \theta < 9$ with an exact confidence coefficient of 0.928.

The straightforward *graphical approach* could be used to simplify the procedure of constructing intervals here. Each of the $m + n$ sample observations is plotted on a graph, the X observations on the abscissa and Y on the ordinate. Then the mn pairings of observations can be easily indicated by dots at all possible intersections. The line $y - x = \theta$ with slope 1 for any number θ divides the pairings into two groups: those on the left and above have $y - x < \theta$, and those on the right and below have $y - x > \theta$. Thus if the rejection region for a size α test is $U \leqslant k$, two lines with slope 1 such that k dots lie on each side of the included band will determine the appropriate values of θ. If the two lines are drawn through the $(k + 1)$st dots from the upper left and lower right, respectively, the values on the vertical axis where these lines have x intercept zero determine the confidence-interval endpoints. In practice, it is often convenient to add or subtract an arbitrary constant from each observation before the pairs are plotted, the number chosen so that all observations are positive and the smallest is close to zero. This does not change the resulting interval for θ, since the parameter θ is invariant under a change in location in both the X and Y populations. This method is illustrated in Fig. 6.1 for the above example where $k = 1$ for $\alpha = 0.072$.

SAMPLE SIZE DETERMINATION

If we are in the process of designing an experiment and specify the size of the test as α and the power of the test as $1 - \beta$, we can determine the sample size required to detect a difference between the populations measured by $p = P(Y > X)$. Noether (1987) showed that an

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

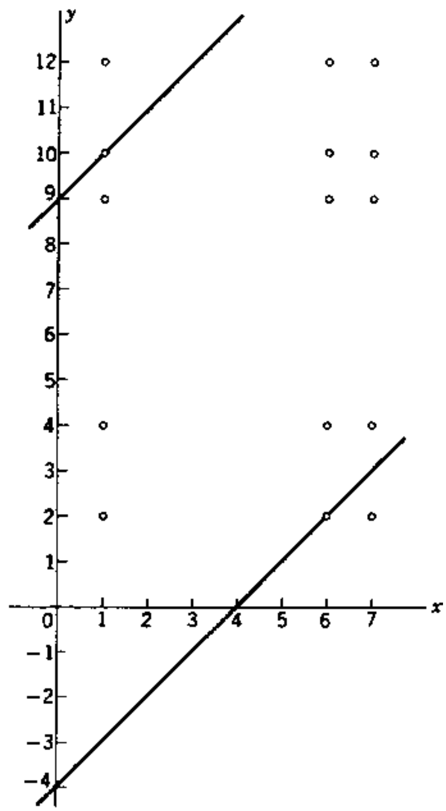


Fig. 6.1 Graphical determination of endpoints.

approximate sample size for a one-sided test based on the Mann-Whitney statistic is

$$N = \frac{(z_\alpha + z_\beta)^2}{12c(1-c)(p-1/2)^2} \tag{6.18}$$

where $c = m/N$ and z_α, z_β are the upper α and β quantiles, respectively, of the standard normal distribution. The corresponding formula for a two-sided test is found by replacing α by $\alpha/2$ in (6.18). Verification of this is left to the reader.

These formulas are based on a normal approximation to the power of the Mann-Whitney test "near" the null hypothesis and can be calculated easily. Note that if we take $c = 0.5$, that is when $m = n$, the formula reduces to the sample size formula (7.15) in Section 5.7 for the Wilcoxon signed-rank test.

As an example, suppose we want to use a one-sided Mann-Whitney test at $\alpha = 0.05$ to detect a difference $p = P(Y > X) = 0.10$ with power 0.90. Suppose we have to take fewer X observations than Y, say $m = 0.8n$. This makes $c = 4/9$. With $z_{0.05} = 1.645$ and $z_{0.10} = 1.28$, we use (6.18) to find

$$N = \frac{(1.645 + 1.28)^2}{12\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)(0.4)} = 18.05$$

Thus we need a total of 19 observations, which translates to $m = 8$, $n = 11$.

Vollandt and Horn (1997) compared Noether's sample size formula to an alternative and found that Noether's approximation is sufficiently reliable with small and large deviations from the null hypothesis.

DISCUSSION

The Mann-Whitney U test is a frequently used nonparametric test that is equivalent to another well-known test, the Wilcoxon rank-sum test, which will be presented independently in Section 8.2. Because the Wilcoxon rank-sum test is easier to use in practice, we postpone giving numerical examples until then. The discussion here applies equally to both tests.

Only independence and continuous distributions need be assumed to test the null hypothesis of identical populations. The test is simple to use for any size samples, and tables of the exact null distribution are widely available. The large-sample approximation is quite adequate for most practical purposes, and corrections for ties can be incorporated in the test statistic. The test has been found to perform particularly well as a test for equal means (or medians), since it is especially sensitive to differences in location. In order to reduce the generality of the null hypothesis in this way, however, we must feel that we can legitimately assume that the populations are identical except possibly for their locations. A particular advantage of the test procedure in this case is that it can be adapted to confidence-interval estimation of the difference in location.

When the populations are assumed to differ only in location, the Mann-Whitney test is directly comparable with Student's t test for means. The asymptotic relative efficiency of U relative to t is *never* less than 0.864, and if the populations are normal, the ARE is quite high at $3/\pi = 0.9550$ (see Chapter 13). The Mann-Whitney test performs better than the t test for some nonnormal distributions. For example, the ARE is 1.50 for the double exponential distribution and 1.09 for the logistic distribution, which are both heavy-tailed distributions.

Many statisticians consider the Mann-Whitney (or equivalently the Wilcoxon rank-sum) test the best nonparametric test for location. Therefore power functions for smaller sample sizes and/or other distributions are of interest. To calculate exact power, we sum the probabilities under the alternative for those arrangements of m X and n Y random variables which are in the rejection region. For any combined arrangement Z where the X random variables occur in the positions $r_1, r_2, \ldots, r_m$ and the Y's in positions $s_1, s_2, \ldots, s_n$, this probability is

$$P(Z) = m!n! \int_{-\infty}^{\infty} \int_{-\infty}^{u_N} \cdots \int_{-\infty}^{u_3} \int_{-\infty}^{u_2} \prod_{i=1}^{m} f_X(u_{r_i}) \prod_{j=1}^{n} f_Y(u_{s_j})\, du_1 \cdots du_N \tag{6.19}$$

which is generally extremely tedious to evaluate. The asymptotic normality of U holds even in the nonnull case, and the mean and variance of U in (6.9) and (6.11) depend only on the parameters p, p_1, and p_2 if the distributions are continuous. Thus, approximations to power can be found if the integrals in (6.3), (6.7), and (6.8) are evaluated. Unfortunately, even under the more specific alternative, $F_Y(x) = F_X(x - \theta)$ for some θ, these integrals depend on both θ and F_X, so that calculating even the approximation to power requires that the basic parent population be specified.

6.7 SUMMARY

In this chapter we have covered four different inference procedures for the null hypothesis that two mutually independent random samples come from identical distributions against the general alternative that they differ in some way. TheWald-Wolfowitz runs test and Kolmogorov-Smirnov two-sample tests are both noted for their generality. Since they are sensitive to any type of difference between the two distributions, they are not very powerful against any specific type of difference that could be stated in the alternative. Their efficiency is very low for location alternatives. They are really useful only in preliminary studies. The median test is primarily sensitive to differences in location and it does have a corresponding confidence interval procedure for estimating the difference in the medians. But it is not very powerful compared to other nonparametric tests for location. Conover, Wehmanen, and Ramsey (1978) examined the power of eight nonparametric tests, including the median test, compared to the *locally most powerful* (LMP) linear rank test when the distribution is exponential for small sample sizes. Even though the median test is asymptotically equivalent to the LMP test, it performed rather poorly. Freidlin and Gastwirth (2000) suggest that the median test "be retired' from routine use" because their simulated power comparisons showed that other tests for location are more powerful for most distributions. Even the Kolmogorov-Smirnov two-sample test was mroe powerful for most of the cases they studied. Gibbons (1964) showed the poor performance of the median test with exact power calculations for small sample sizes. Further, the hand calculations for an exact median test based on the hypergeometric distribution are quite tedious even for small sample sizes. The median test continues to be of theoretical interest, however, because it is valid under such weak conditions and has interesting theoretical properties.

The Mann-Whitney test is far preferable as a test of location for general use, as are the other rank tests for location to be covered later in Chapter 8. The Mann-Whitney test also has a corresponding procedure for confidence interval estimation of the difference in population medians. And we can estimate the sample size needed to carry out a test at level α to detect a stated difference in locations with power $1-\beta$.

PROBLEMS

6.1. Use the graphical method of Hodges to find $P(D^+_{m,n} \geqslant d)$, where d is the observed value of $D^+_{m,n} = \max_x[S_m(x) - S_n(x)]$ in the arrangement *xyyxyx*.

6.2. For the median-test statistic derive the complete null distribution of U for $m = 6, n = 7$, and set up one- and two-sided critical regions when $\alpha = 0.01, 0.05$, and 0.10.

6.3. Find the large-sample approximation to the power function of a two-sided median test for $m = 6, n = 7, \alpha = 0.10$, when F_X is the standard normal distribution.

6.4. Use the recursion relation for the Mann-Whitney test statistic given in (6.14) to generate the complete null probability distribution of U for all $m + n \leqslant 4$.

6.5. Verify the expressions given in (6.15) for the moments of U under H_0.

6.6. Answer parts (a) to (c) using (i) the median-test procedure and (ii) the Mann-Whitney test procedure (use tables) for the following two independent random samples drawn from continuous populations which have the same form but possibly a difference of θ in their locations:

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

X	79	13	138	129	59	76	75	53
Y	96	141	133	107	102	129	110	104

(a) Using the significance level 0.10, test

H_0: $\theta = 0$ versus H_1: $\theta \neq 0$

(b) Give the exact level of the test in (a).

(c) Give a confidence interval on θ, with an exact confidence coefficient corresponding to the exact level noted in (b).

6.7. Represent a sample of m X and n Y random variable by a path of $m + n$ steps, the ith step being one unit up or to the right according as the ith from the smallest observation in the combined sample is an X or a Y, respectively. What is the algebraic relation between the area under the path and the Mann-Whitney statistic?

6.8. Can you think of other functions of the difference $S_m(x) - S_n(x)$ (besides the maximum) which could also be used for distribution-free tests of the equality of two population distributions?

6.9. The 2000 census statistics for Alabama give the percentage changes in population between 1990 and 2000 for each of the 67 counties. These counties were divided into two mutually independent groups, rural and nonrural, according to population size of less than 25,000 in 2000 or not. Random samples of nine rural and seven nonrural counties gave the following data on percentage population change:

Rural	1.1,−21.7,−16.3, −11.3,−10.4,−7.0,−2.0,1.9,6.2
Nonrural	−2.4, 9.9, 14.2, 18.4, 20.1, 23.1, 70.4

Use all of the methods of this chapter to test the null hypothesis of equal distributions. (*Hint*: Notice that $m \neq n$ here.)

6.10. (a) Show that the distribution of the precedence statistic $P_{(i)}$ under the null hypothesis ($F_X = F_Y$), given in Problem 2.28(c), can be expressed as

$$P(P_{(i)} = j \mid H_0) = \frac{n}{m+n} \frac{\binom{m}{j}\binom{n-1}{i-1}}{\binom{m+n-1}{j+i-1}}$$

$$= \frac{i}{j+1} \frac{\binom{m}{j}\binom{n}{i}}{\binom{m+n}{j+i}} \qquad j = 0, 1, \ldots, m$$

These relationships are useful in calculating the null distribution of precedence statistics using tables of a hypergeometric distribution.

(b) Hence show that the null distribution of the control median test statistic V, with $n = 2r + 1$, can be expressed as

$$\frac{2r+1}{m+2r+1} \frac{\binom{m}{j}\binom{2r}{r}}{\binom{m+2r}{j+r}} \qquad j = 0, 1, \ldots, m$$

(c) Prepare a table of the cumulative probabilities of the null distribution of V for some suitable values of m and n (odd).

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

6.11. For the control median test statistic V, use Problem 2.28, or otherwise, to show that when $F_X = F_Y$,

$$E(V) = \frac{m}{2} \qquad \text{and} \qquad \text{var}(V) = \frac{2r+m+2}{4m(2r+3)}$$

[*Hint*: Use the fact that $E(X) = E_Y E(X|Y)$ and $\text{var}(X) = \text{var}_Y E(X|Y) + E_Y \text{var}(X|Y)$]

6.12. Show that when $m, n \to \infty$ such that $m/(m+n) \to \lambda$, $0 < \lambda < 1$, then the null distribution of the precedence statistic $P_{(i)}$ given in Problem 6.10 tends to the negative binomial distribution with parameters i and λ, or

$$\binom{j+i-1}{i-1}\lambda^j(1-\lambda)^j \qquad j = 0, 1, \ldots, m \qquad \text{(Sen, 1964)}$$

6.13. In some applications the quantity $\xi_p = F_X(\kappa_p)$, where κ_p is the pth quantile of F_Y, is of interest. Let $\lim_{n\to\infty}(m/n) = \lambda$, where λ is a fixed quantity, and let $\{r_n\}$ be a sequence of positive integers such that $\lim_{n\to\infty}(r_n/n) = p$. Finally let $V_{m,n}$ be the number of X observations that do not exceed $Y_{(r_n)}$.

(a) Show that $m^{-1}V_{m,n}$ is a consistent estimator of ξ_p.

(b) Show that the random variable $m^{1/2}[m^{-1}V_{m,n} - \xi_p]$ is asymptotically normally distributed with mean zero and variance

$$\xi_p(1-\xi_p) + \lambda_p(1-p)\frac{f_X^2(\kappa_p)}{f_Y^2(\kappa_p)}$$

where f_X and f_Y are the density functions corresponding to F_X and F_Y, respectively.
(Gastwirth, 1968; Chakraborti and Mukerjee; 1990)

6.14. A sample of three girls and five boys are given instructions on how to complete a certain task. Then they are asked to perform the task over and over until they complete it correctly. The number of repetitions necessary for correct completion are 1, 2, and 5 for the girls and 4, 8, 9, 10, and 12 for the boys. Find the P value for the alternative that on the average the girls learn the task faster than the boys, and find a confidence interval estimate for the difference $\theta = M_Y - M_X$ with a confidence coefficient at least equal to 0.85, using the median test.

6.15. A researcher is interested in learning if a new drug is better than a placebo in treating a certain disease. Because of the nature of the disease, only a limited number of patients can be found. Out of these, 5 are randomly assigned to the placebo and 5 to the new drug. Suppose that the concentration of a certain chemical in blood is measured and smaller measurements are better. The data are as follows:

$$\text{Drug: } 3.2, 2.1, 2.3, 1.2, 1.5 \qquad \text{Placebo: } 3.4, 3.5, 4.1, 1.7, 2.1$$

(a) Use the median test and the control median test to test the hypothesis. For each test give the null hypothesis, the alternative hypothesis, the value of the test statistic, the exact and the approximate P value and a conclusion. What assumptions are we making?

(b) Use the median test to calculate a confidence interval for the difference between the medians. What is the highest possible level of confidence? What assumptions are we making for this procedure?

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

7

Linear Rank Statistics and the General Two-Sample Problem

7.1 INTRODUCTION

The general two-sample problem was described in Chapter 6 and some tests were presented which were all based on various criteria related to the combined ordered arrangement of the two sets of sample observations. Many statistical procedures applicable to the two-sample problem are based on the rank-order statistics for the combined samples, since various functions of these rank-order statistics can provide information about the possible differences between populations. For example, if the X population has a larger mean than the Y population, the sample values will reflect this difference if most of the ranks of the X values exceed the ranks of the Y values.

Many commonly used two-sample rank tests can be classified as linear combinations of certain indicator variables for the combined ordered samples. Such functions are often called *linear rank statistics*. This unifying concept will be defined in the next section, and then

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

some of the general theory of these linear rank statistics will be presented. Particular linear rank tests will then be treated in Chapters 8 and 9 for the location and scale problems respectively.

7.2 DEFINITION OF LINEAR RANK STATISTICS

Assume we have two independent random samples, $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ drawn from populations with continuous cumulative distribution functions F_X and F_Y, respectively. Under the null hypothesis

$$H_0\colon F_X(x) = F_Y(x) = F(x) \qquad \text{for all } x, F \text{ unspecified}$$

we then have a single set of $m + n = N$ random observations from the common but unknown population, to which the integer ranks $1, 2, \ldots, N$ can be assigned.

In accordance with the definition for the rank of an observation in a single sample given in (5.1) of Section 5.5, a functional definition of the rank of an observation in the combined sample with no ties can be given as

$$\begin{aligned} r_{XY}(x_i) &= \sum_{k=1}^{m} S(x_i - x_k) + \sum_{k=1}^{n} S(x_i - y_k) \\ r_{XY}(y_i) &= \sum_{k=1}^{m} S(y_i - x_k) + \sum_{k=1}^{n} S(y_i - y_k) \end{aligned} \tag{2.1}$$

where

$$s(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u \geqslant 0 \end{cases}$$

However, it is easier to denote the combined ordered sample by a vector of indicator random variables as follows. Let

$$Z = (Z_1, Z_2, \ldots, Z_N)$$

where $Z_i = 1$ if the ith random variable in the combined ordered sample is an X and $Z_i = 0$ if it is a Y, for $1, 2, \ldots, N$, with $N = m + n$. The rank of the observation for which Z_i is an indicator is i, and therefore the vector Z indicates the rank-order statistics of the combined samples and in addition identifies the sample to which each observation belongs.

For example, given the observations $(X_1, X_2, X_3, X_4) = (2,9,3,4)$ and $(Y_1, Y_2, Y_3) = (1,6,10)$, the combined ordered sample is (1,2,3, 4,6,9,10) or $(Y_1, X_1, X_3, X_4, Y_2, X_2, Y_3)$, and the corresponding Z vector is (0,1,1,1,0,1,0). Since $Z_6 = 1$, for example, an X observation (in particular X_2) has rank 6 in the combined ordered array.

Many of the statistics based on rank-order statistics which are useful in the two-sample problem can be easily expressed in terms of this notation. An important class of statistics of this type is called a *linear rank statistic*, defined as a linear function of the indicator variables Z, as

$$T_N(Z) = \sum_{i=1}^{N} a_i Z_i \tag{2.2}$$

where the a_i are given constants called weights or scores. It should be noted that the statistic T_N is linear in the indicator variables and no similar restriction is implied for the constants.

7.3 DISTRIBUTION PROPERTIES OF LINEAR RANK STATISTICS

We shall now prove some general properties of T_N in order to facilitate the study of particular linear-rank-statistic tests later.

Theorem 3.1 *Under the null hypothesis H_0: $F_X(x) = F_Y(x) = F(x)$ for all x, we have for all $i = 1, 2, \ldots, N$,*

$$E(Z_i) = \frac{m}{N} \qquad \operatorname{var}(Z_i) = \frac{mn}{N^2} \qquad \operatorname{cov}(Z_i, Z_j) = \frac{-mn}{N^2(N-1)} \tag{3.1}$$

Proof Since

$$f_{Z_i(z_i)} = \begin{cases} m/N & \text{if } z_i = 1 \\ n/N & \text{if } z_i = 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i = 1, 2, \ldots, N$$

is the Bernoulli distribution, the mean and variance are

$$E(Z_i) = m/N \qquad \operatorname{var}(Z_i) = mn/N^2$$

For the joint moments, we have for $i \neq j$,

$$E(Z_i Z_j) = P(Z_i = 1 \cap Z_j = 1) = \frac{\binom{m}{2}}{\binom{N}{2}} = \frac{m(m-1)}{N(N-1)}$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

so that

$$\operatorname{cov}(Z_i, Z_j) = \frac{m(m-1)}{N(N-1)} - \left(\frac{m}{N}\right)^2 = \frac{-mn}{N^2(N-1)}$$

Theorem 3.2 *Under the null hypothesis $H_0\colon F_X(x) = F_Y(x) = F(x)$ for all x,*

$$E(T_N) = m\sum_{i=1}^{N}\frac{a_i}{N}$$

$$\operatorname{var}(T_N) = \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}a_i^2 - \left(\sum_{i=1}^{N}a_i\right)^2\right] \tag{3.2}$$

Proof

$$E(T_N) = \sum_{i=1}^{N}a_iE(Z_i) = m\sum_{i=1}^{N}\frac{a_i}{N}$$

$$\begin{aligned}\operatorname{var}(T_N) &= \sum_{i=1}^{N}a_i^2\operatorname{var}(Z_i) + \sum_{i\neq j}\sum a_ia_j\operatorname{cov}(Z_i, Z_j)\\ &= \frac{mn\sum_{i=1}^{N}a_i^2}{N^2} - \frac{mn\sum_{i\neq j}\sum a_ia_j}{N^2(N-1)}\\ &= \frac{mn}{N^2(N-1)}\left(N\sum_{i=1}^{N}a_i^2 - \sum_{i=1}^{N}a_i^2 - \sum_{i\neq j}\sum a_ia_j\right)\\ &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}a_i^2 - \left(\sum_{i=1}^{N}a_i\right)^2\right]\end{aligned}$$

Theorem 3.3 *If $B_N = \sum_{i=1}^{N}b_iZ_i$ and $T_N = \sum_{i=1}^{N}a_iZ_i$ are two linear rank statistics, under the null hypothesis $H_0\colon F_X(x) = F_Y(x) = F(x)$ for all x,*

$$\operatorname{cov}(B_N, T_N) = \frac{mn}{N^2(N-1)}\left(N\sum_{i=1}^{N}a_ib_i - \sum_{i=1}^{N}a_i\sum_{i=1}^{N}b_i\right)$$

Proof

$$\begin{aligned}\operatorname{cov}(B_N, T_N) &= \sum_{i=1}^{N}a_ib_i\operatorname{var}(Z_i) + \sum_{i\neq j}\sum a_ib_j\operatorname{cov}(Z_i, Z_j)\\ &= \frac{mn}{N^2}\sum_{i=1}^{N}a_ib_i - \frac{mn}{N^2(N-1)}\sum_{i\neq j}\sum a_ib_i\end{aligned}$$

$$= \frac{mn}{N^2(N-1)} \left(N \sum_{i=1}^{N} a_i b_i - \sum_{i=1}^{N} a_i b_i - \sum \sum_{i \neq j} a_i b_j \right)$$
$$= \frac{mn}{N^2(N-1)} \left(N \sum_{i=1}^{N} a_i b_i - \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \right)$$

Using these theorems, the exact moments under the null hypothesis can be found for any linear rank statistics. The exact null probability distribution of T_N depends on the probability distribution of the vector Z, which indicates the ranks of the X and Y random variables. This distribution was given in Eq. (6.19) of Section 6.6 for any distributions F_X and F_Y. In the null case, $F_X = F_Y = F$, say, and the equation reduces to

$$P(Z) = m!n! \int_{-\infty}^{\infty} \int_{-\infty}^{u_N} \cdots \int_{-\infty}^{u_2} \prod_{i=1}^{m} f(u_{r_i}) \prod_{j=1}^{n} f(u_{s_j})\, du_1 \cdots du_N$$

where $r_1, r_2, \ldots, r_m$ and $s_1, s_2, \ldots, s_n$ are the ranks of the X and Y random variables, respectively, in the arrangement Z. Since the distributions are identical, the product in the integrand is the same for all subscripts, or

$$P(Z) = m!n! \int_{-\infty}^{\infty} \int_{-\infty}^{u_N} \cdots \int_{-\infty}^{u_2} \prod_{i=1}^{N} f(u_i) du_1 \cdots du_N$$
$$= \frac{m!n!}{N!} \tag{3.3}$$

The final result follows from the fact that except for the terms $m!n!$, $P(Z)$ is the integral over the entire region of the density function of the N order statistics for a random sample from the population F. Since $\binom{m+n}{m} = \binom{N}{m}$ is the total number of distinguishable Z vectors, i.e., distinguishable arrangements of m ones and n zeros, the result in (3.3) implies that all vectors Z are equally likely under H_0.

Since each Z occurs with probability $1/\binom{N}{m}$, the exact probability distribution under the null hypothesis of any linear rank statistic can always be found by direct enumeration. The values of $T_N(Z)$ are calculated for each Z, and the probability of a particular value k is the number of Z vectors which lead to that number k divided by $\binom{N}{m}$. In other words, we have

$$P[T_N(Z) = k] = t(k) \Big/ \binom{N}{m} \tag{3.4}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where $t(k)$ is the number of arrangements of m X and n Y random variables such that $T_N(Z) = k$. Naturally, the tediousness of enumeration increases rapidly as m and n increase. For some statistics, recursive methods are possible. STATXACT calculates the exact P value for a linear rank test based on a complete enumeration of the values of the test statistic. Here the data are permuted (rearranged) in all possible ways under the null hypothesis that is being tested. The value of the test statistic is calculated for each permutation of the data; these values constitute the *permutation distribution* and allow calculation of the exact P value for any test based on ranks of any set of data.

When the null distribution of a linear rank statistic is known to be symmetric, only one-half of the distribution needs to be generated. The statistic is symmetric about its mean μ if for every $k \neq 0$,

$$P[T_N(Z) - \mu = k] = P[T_N(Z) - \mu = -k]$$

Suppose that for every vector Z of m ones and n zeros, a conjugate vector Z' of m ones and n zeros exists such that whenever $T_N(Z) = \mu + k$, we have $T_N(Z') = \mu - k$. Then the frequency of the number $\mu + k$ is the same as that of $\mu - k$, and the distribution is symmetric. The condition for symmetry of a linear rank statistic then is that

$$T_N(Z) + T_N(Z') = 2\mu$$

The following theorem establishes a simple relation between the scores which will ensure the symmetry of $T_N(Z)$.

Theorem 3.4 *The null distribution of $T_N(Z)$ is symmetric about its mean $\mu = m \sum_{i=1}^{N} a_i/N$ whenever the weights satisfy the relation*

$$a_i + a_{N-i+1} = c \qquad c = \text{constant for } i = 1, 2, \ldots, N$$

Proof For any vector $Z = (Z_1, Z_2, \ldots, Z_N)$ of m ones and n zeros, define the conjugate vector $Z' = (Z'_1, Z'_2, \ldots, Z'_N)$, where $Z'_i = Z_{N-i+1}$. Then

$$\begin{aligned} T_N(Z) + T_N(Z') &= \sum_{i=1}^{N} a_i Z_i + \sum_{i=1}^{N} a_i Z_{N-i+1} \\ &= \sum_{i=1}^{N} a_i Z_i + \sum_{j=1}^{N} a_{N-j+1} Z_j \\ &= \sum_{i=1}^{N} (a_i + a_{N-i+1}) Z_i = c \sum_{i=1}^{N} Z_i = cm \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Since $E[T_N(Z)] = E[T_N(Z')]$, we must have $cm = 2\mu$, or $c = 2\mu/m = 2\sum_{i=1}^{N} a_i/N$.

The next theorem establishes the symmetry of *any* linear rank statistic when $m = n$.

Theorem 3.5 *The null distribution of $T_N(Z)$ is symmetric about its mean for any set of weights if $m = n = N/2$.*

Proof Since $m = n$, we can define our conjugate Z' with ith component $Z'_i = 1 - Z_i$. Then

$$T_N(Z) + T_N(Z') = \sum_{i=1}^{N} a_i Z_i + \sum_{i=1}^{N} a_i(1 - Z_i) = \sum_{i=1}^{N} a_i = 2\mu$$

A rather special but useful case of symmetry is given as follows.

Theorem 3.6 *The null distribution of $T_N(Z)$ is symmetric about its mean μ if N is even and the weights are $a_i = i$ for $i \leqslant N/2$ and $a_i = N - i + 1$ for $i > N/2$.*

Proof The appropriate conjugate Z' has components $Z'_i = Z_{i+N/2}$ for $i \leqslant N/2$ and $Z'_i = Z_{i-N/2}$ for $i > N/2$. Then

$$\begin{aligned}
T_N(Z) + T_N(Z') &= \sum_{i=1}^{N/2} iZ_i + \sum_{i=N/2+1}^{N} (N - i + 1)Z_i \\
&\quad + \sum_{i=1}^{N/2} iZ_{N/2+i} + \sum_{i=N/2+1}^{N} (N - i + 1)Z_{i-N/2} \\
&= \sum_{i=1}^{N/2} iZ_i + \sum_{i=N/2+1}^{N} (N - i + 1)Z_i \\
&\quad + \sum_{j=N/2+1}^{N} \left(j - \frac{N}{2}\right)Z_j + \sum_{j=1}^{N/2} \left(\frac{N}{2} - j + 1\right)Z_j \\
&= \sum_{i=1}^{N/2} \left(\frac{N}{2} + 1\right)Z_i + \sum_{i=N/2+1}^{N} \left(\frac{N}{2} + 1\right)Z_i \\
&= m\left(\frac{N}{2} + 1\right) = 2\mu
\end{aligned}$$

In determining the frequency $t(k)$ for any value k which is assumed by a linear-rank-test statistic, the number of calculations

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

required may be reduced considerably by the following properties of $T_N(Z)$, which are easily verified.

Theorem 3.7

Property 1: *Let*

$$T = \sum_{i=1}^{N} a_i Z_i \qquad \text{and} \qquad T' = \sum_{i=1}^{N} a_i Z_{N-i+1}$$

Then

$$T = T' \qquad \text{if } a_i = a_{N-i+1} \quad \text{for } i = 1, 2, \ldots, N$$

Property 2: *Let*

$$T = \sum_{i=1}^{N} a_i Z_i \qquad \text{and} \qquad T' = \sum_{i=1}^{N} a_i (1 - Z_i)$$

Then

$$T + T' = \sum_{i=1}^{N} a_i$$

Property 3: *Let*

$$T = \sum_{i=1}^{N} a_i Z_i \qquad \text{and} \qquad T' = \sum_{i=1}^{N} a_i (1 - Z_{N-i+1})$$

Then

$$T + T' = \sum_{i=1}^{N} a_i \qquad \text{if } a_i = a_{N-i+1} \text{ for } i = 1, 2, \ldots, N$$

For large samples, that is, $m \to \infty$ and $n \to \infty$ in such a way that m/n remains constant, an approximation exists which is applicable to the distribution of almost all linear rank statistics. Since T_N is a linear combination of the Z_i, which are identically distributed (though dependent) random variables, a generalization of the central-limit theorem allows us to conclude that the probability distribution of a standardized linear rank statistic $T_N - E(T_N)/\sigma(T_N)$ approaches the standard normal probability distribution subject to certain regularity conditions.

The foregoing properties of a linear rank statistic hold only in the hypothesized case of identical populations. Chernoff and Savage

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

(1958) have proved that the asymptotic normality property is valid also in the nonnull case, subject to certain regularity conditions relating mainly to the smoothness and size of the weights. The expressions for the mean and variance will be given here, since they are also useful in investigating consistency and efficiency properties of most two-sample linear rank statistics.

A key feature in the Chernoff-Savage theory is that a linear rank statistic can be represented in the form of a Stieltjes integral. Thus, if the weights for a linear rank statistic are functions of the ranks, an equivalent representation of $T_N = \sum_{i=1}^{N} a_i Z_i$ is

$$T_N = m \int_{-\infty}^{\infty} J_N[H_N(x)]\, dS_m(x)$$

where the notation is defined as follows:

1. $S_m(x)$ and $S_n(x)$ are the empirical distribution functions of the X and Y samples, respectively.
2. $m/N \to \lambda_N,\ 0 < \lambda_N < 1$.
3. $H_N(x) = \lambda_N S_m(x) + (1 - \lambda_N) S_n(x)$, so that $H_N(x)$ is the proportion of observations from either sample which do not exceed the value x, or the empirical distribution function of the combined sample.
4. $J_N(i/N) = a_i$.

This Stieltjes integral form is given here because it appears frequently in the journal literature and is useful for proving theoretical properties. Since the following theorems are given here without proof anyway, the student not familiar with Stieltjes integrals can consider the following equivalent representation:

$$T'_N = m \sum_{\substack{\text{over all x such} \\ \text{that } p(x)>0}} J_N[H_N(x)] p(x)$$

where

$$p(x) = \begin{cases} 1/m & \text{if } x \text{ is the observed value of an } X \text{ random variable} \\ 0 & \text{otherwise} \end{cases}$$

For example, in the simplest case where $a_i = i/N, J_N[H_N(x)] = H_N(x)$ and

$$T_N = m \int_{-\infty}^{\infty} H_N(x) dS_m(x) = \frac{m}{N} \int_{-\infty}^{\infty} [m S_m(x) + n S_n(x)]\, dS_m(x)$$

$$= \frac{m}{N} \int_{-\infty}^{\infty} (\text{number of observations in the combined sample} \leqslant x)$$
$$\times (1/m \text{ if } x \text{ is the value of an } X \text{ random variable and } 0 \text{ otherwise})$$
$$= \frac{1}{N} \sum_{i=1}^{N} iZ_i$$

Now when the X and Y samples are drawn from the continuous populations F_X and F_Y, respectively, we define the combined population cdf as

$$H(x) = \lambda_N F_X(x) + (1 - \lambda_N) F_Y(x)$$

The Chernoff and Savage theorem stated below is subject to certain regularity conditions not explicitly stated here, but given in Chernoff and Savage (1958).

Theorem 3.8 *Subject to certain regularity conditions, the most important of which are that* $J(H) = \lim_{N\to\infty} J_N(H)$,

$$\left|J^{(r)}(H)\right| = |d^r J(H)/dH^r| \leqslant K|H(1-H)|^{-r-1/2+\delta}$$
for $r = 0, 1, 2$ *and some* $\delta > 0$ *and* K *any constant which does not depend on* $m, n, N, F_X,$ or F_Y,

then for λ_N *fixed,*

$$\lim_{N\to\infty} P\left(\frac{T_N/m - \mu_N}{\sigma_N} \leqslant t\right) = \Phi(t)$$

where

$$\mu_N = \int_{-\infty}^{\infty} J[H(x)] f_X(x)\, dx$$

$$N\sigma_N^2 = 2\frac{1-\lambda_N}{\lambda_N}\left\{\lambda_N \iint_{-\infty<x<y<\infty} F_Y(x)[1 - F_Y(y)] J'[H(x)] J'[H(y)]\right.$$
$$\times f_X(x) f_Y(y)\, dx\, dy + (1-\lambda_N) \iint_{-\infty<x<y<\infty} F_X(x)[1 - F_X(y)]$$

$$\left.\times J'[H(x)]J'[H(y)]\,f_X(x)f_Y(y)\,dx\,dy\right\}$$

provided $\sigma_N \neq 0$.

Corollary 3.8 *If X and Y are identically distributed with common distribution $F(x) = F_X(x) = F_Y(x)$, we have*

$$\mu_N = \int_0^1 J(u)\,du$$

$$\begin{aligned}
N\lambda_N\sigma_N^2 &= 2(1-\lambda_N)\iint\limits_{0<x<y<1} x(1-y)J'(x)J'(y)\,dx\,dy \\
&= 2(1-\lambda_N)\iiiint\limits_{0<u<x<y<v<1} J'(x)J'(y)\,dx\,dy\,du\,dv \\
&= 2(1-\lambda_N)\iint\limits_{0<u<v<1}\int_u^v\int_x^v J'(x)J'(y)\,dy\,dx\,du\,dv \\
&= 2(1-\lambda_N)\iint\limits_{0<u<v<1}\int_u^v [J(v)-J(x)]J'(x)\,dx\,du\,dv \\
&= 2(1-\lambda_N)\iint\limits_{0<u<v<1}\left[J(v)J(x)-\frac{J^2(x)}{2}\right]\Bigg\|_u^v du\,dv \\
&= (1-\lambda_N)\iint\limits_{0<u<v<1}[J^2(v)-2J(v)J(u)+J^2(u)]\,du\,dv \\
&= (1-\lambda_N)\left[\int_0^1 vJ^2(v)\,dv + \int_0^1(1-u)J^2(u)du \right.\\
&\qquad \left. - \int_0^1 J(u)\,du\int_0^1 J(v)\,dv\right] \\
&= (1-\lambda_N)\left\{\int_0^1 J^2(u)\,du - \left[\int_0^1 J(u)\,du\right]^2\right\}
\end{aligned}$$

These expressions are equivalent to those given in Theorem 3.2 for $a_i = J_N(i/N)$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

7.4 USEFULNESS IN INFERENCE

The general alternative to the null hypothesis in the two-sample problem is simply that the populations are not identical, i.e.,

$$F_X(x) \neq F_Y(x) \qquad \text{for some } x$$

or the analogous one-sided general alternative, which states a directional inequality such as

$$F_X(x) \leqslant F_Y(x) \qquad \text{for all } x$$

The two-sample tests considered in Chapter 6, namely the Kolmogorov-Smirnov, Wald-Wolfowitz runs, Mann-Whitney, and median tests, are all appropriate for these alternatives. In most parametric two-sample situations, the alternatives are much more specific, as in the t and F tests for comparison of means and variances, respectively. Although all of the two-sample rank tests are for the same null hypothesis, particular test statistics may be especially sensitive to a particular form of alternative, thus increasing their power against that type of alternative.

Since any set of scores $a_1, a_2, \ldots, a_N$ may be employed for the coefficients in a linear rank statistic, this form of test statistic lends itself particularly well to more specific types of alternatives the user might have in mind. The appropriateness of choice depends on the type of difference between populations one hopes to detect. The simplest type of situation to deal with is where the statistician has enough information about the populations to feel that if a difference exists, it is only in location or only in scale. These will be called, respectively, the *two-sample location problem* and the *two-sample scale problem*. In the following two chapters we shall discuss briefly some of the better-known and more widely accepted linear rank statistics useful in these problems. No attempt will be made to provide recommendations regarding which to use. The very generality of linear rank tests makes it difficult to make direct comparisons of power functions, since calculation of power requires more specification of the alternative probability distributions and moments. A particular test might have high power against normal alternatives but perform poorly for the gamma distribution. Furthermore, calculation of the exact power of rank tests is usually quite difficult. We must be able to determine the probability distribution of the statistic $T_N(Z)$ or the arrangement Z as in Eq. (6.19) of Section 6.6 under the specified alternative and sum these probabilities over those arrangements Z in the rejection region specified by the test. STATXACT can be useful in calculating the exact power of linear rank tests. Isolated and specific comparisons of power between nonparametric tests have received much attention in the literature, and the reader is referred to Savage's *Bibliography* (1962) for some early references. However, calculation of asymptotic relative efficiency of linear rank tests versus the t and F tests for normal alternatives is not particularly difficult. Therefore, information regarding the ARE's of the tests presented here for the location and scale problems will be provided.

PROBLEMS

7.1. One of the simplest linear rank statistics is defined as

$$W_N = \sum_{i=1}^{N} iZ_i$$

This is the Wilcoxon rank-sum statistic to be discussed on the next chapter. Use Theorem 3.2 to evaluate the mean and variance of W_N.

7.2. Express the two-sample median-test statistic U defined in Section 6.4 in the form of a linear rank statistic and use Theorem 3.2 to find its mean and variance. *Hint*: For the appropriate argument k, use the functions $S(k)$ defined as for (2.1).

7.3. Prove the three properties stated in Theorem 3.7.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

8
Linear Rank Tests for the Location Problem

8.1 INTRODUCTION

Suppose that two independent samples of sizes m and n are drawn from two continuous populations so that we have $N = m + n$ observations in total. We wish to test the null hypothesis of identical distributions. The *location alternative* is that the populations are of the same form but with a different measure of central tendency. This can be expressed symbolically as follows:

$$H_0: F_Y(x) = F_X(x) \qquad \text{for all } x$$
$$H_L: F_Y(x) = F_X(x - \theta) \qquad \text{for all } x \text{ and some } \theta \neq 0$$

The cumulative distribution of the Y population under H_L is functionally the same as that of the X population but shifted to the left if $\theta < 0$ and shifted to the right if $\theta > 0$, as shown in Figure 1.1.

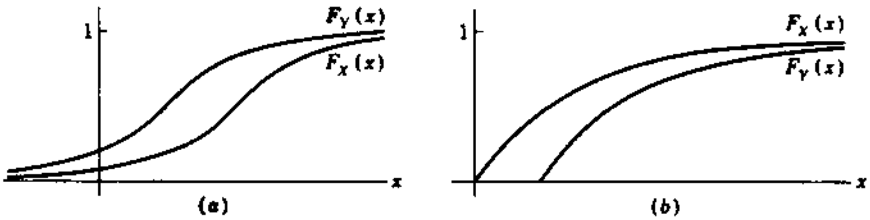


Fig. 1.1 $F_Y(x) = F_X(x - \theta)$. (a) F_X normal, $\theta < 0$; (b) F_X exponential, $\theta > 0$.

Therefore the Y's are stochastically larger than the X's when $\theta > 0$ and the Y's are stochastically smaller than the X's when $\theta < 0$. Thus, when $\theta < 0$, for example, the median of the X population is larger than the median of the Y population.

If it is reasonable to assume that F_X is the cumulative normal distribution, then the mean and median coincide and a one-sided normal-theory test with equal but unknown variances of the hypothesis

$$\mu_Y - \mu_X = 0 \qquad \text{versus} \qquad \mu_Y - \mu_X < 0$$

is equivalent to the general location alternative with $\theta = \mu_Y - \mu_X < 0$. The best parametric test against this alternative is the t statistic with $m + n - 2$ degrees of freedom:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}}\sqrt{\dfrac{m+n}{mn}}} \tag{1.1}$$

The t test statistic has been shown to be robust under the assumptions of normality and equal variances. However, there are many good and simple nonparametric tests for the location problem which do not require specification of the underlying population, such as assuming normality. Many of these are based on ranks since the ranks of the X's relative to the ranks of the Y's provide information about the relative size of the population medians. In the form of a linear rank statistic, any set of scores which are nondecreasing or nonincreasing in magnitude would allow the statistic to reflect a combined ordered sample in which most of the X's are larger than the Y's, or vice versa. The Wilcoxon rank-sum test is one of the best known and easiest to use, since it employs scores which are positive integers. The other tests which will be covered in this chapter are the Terry-Hoeffding-normal-scores test, inverse-normal-scores test, and percentile modified rank tests. There are many others discussed in the literature.

8.2 THE WILCOXON RANK-SUM TEST

The ranks of the X's in the combined ordered arrangement of the two samples would generally be larger than the ranks of the Y's if the median of the X population exceeds the median of the Y population. Therefore, Wilcoxon (1945) proposed a test where we accept the one-sided location alternative $H_L: \theta < 0$ ($X \overset{\text{ST}}{>}$ Y) if the sum of the ranks of the X's is too large or $H_L: \theta > 0$ ($X \overset{\text{ST}}{<} Y$) if the sum of the ranks of the X's is too small and the two-sided location alternative $H_L: \theta \neq 0$ if the sum of the ranks of the X's is either too large or too small. This function of the ranks expressed as a linear rank statistic has the simple weights $a_i = i,\ i = 1, 2, \ldots, N$. In other words, the *Wilcoxon rank-sum test* statistic is

$$W_N = \sum_{i=1}^{N} iZ_i$$

where the Z_i are the indicator random variables as defined for (7.2.2) [Eq. (2.2) of Section 7.2].

If there are no ties, the exact mean and variance of W_N under the null hypothesis of equal distributions are easily found from Theorem 7.3.2 to be

$$E(W_N) = \frac{m(N+1)}{2} \qquad \text{var}(W_N) = \frac{mn(N+1)}{12}$$

Verification of these facts is left for the reader. If $m \leq n$, the value of W_N has a minimum of $\sum_{i=1}^{m} i = m(m+1)/2$ and a maximum of $\sum_{i=N-m+1}^{N} i = (2N - m + 1)/2$. Further, from Theorem 7.3.4, since

$$a_i + a_{N-i+1} = N + 1 \qquad \text{for } i = 1, 2, \ldots, N$$

the statistic is symmetric about its mean. The exact null probability distribution can be obtained systematically by enumeration using these properties. For example, suppose $m = 3,\ n = 4$. There are $\binom{7}{3} = 35$ possible distinguishable configurations of 1's and 0's in the vector Z, but these need not be enumerated individually. W_N will range between 6 and 18, symmetric about 12, the values occurring in conjunction with the ranks in Table 2.1, from which the complete

Table 2.1 Distribution of W_N

Value of W_N	*Ranks of X's*	*Frequency*
18	5,6,7	1
17	4,6,7	1
16	3,6,7;4,5,7	2
15	2,6,7;3,5,7;4,5,6	3
14	1,6,7;2,5,7;3,4,7;3,5,6	4
13	1,5,7;2,4,7;2,5,6;3,4,6	4
12	1,4,7;2,3,7;1,5,6;2,4,6;3,4,5	5

probability distribution is easily found. For example, Table 2.1 shows that $P(W_N \geqslant 17) = 2/35 = 0.0571$.

Several recursive schemes are also available for generation of the distribution. The simplest to understand is analogous to the recursion relations given in (5.7.8) for the Wilcoxon signed-rank statistic and (6.6.14) for the Mann-Whitney statistic. If $r_{m,n}(k)$ denotes the number of arrangements of m X and n Y random variables such that the sum of X ranks is equal to k, it is evident that

$$r_{m,n}(k) = r_{m-1,n}(k-N) + r_{m,n-1}(k)$$

and

$$f_{W_N}(k) = p_{m,n}(k) = \left[r_{m-1,n}(k-N) + r_{m,n-1}(k)\right] \Big/ \binom{m+n}{m}$$

or

$$(m+n)p_{m,n}(k) = mp_{m-1,n}(k-N) + np_{m,n-1}(k) \tag{2.2}$$

Tail probabilities for the null distribution of the Wilcoxon rank-sum test statistic are given in Table J of the Appendix for $m \leqslant n \leqslant 10$. More extensive tables are available in Wilcoxon, Katti, and Wilcox (1972).

For larger sample sizes, generation of the exact probability distribution is rather time-consuming. However, the normal approximation to the distribution or rejection regions can be used because of the asymptotic normality of the general linear rank statistic (Theorem 7.3.8). The normal approximation for W_N has been shown to be accurate enough for most practical purposes for combined sample sizes N as small as 12.

The midrank method is easily applied to handle the problem of ties. The presence of a moderate number of tied observations seems to have little effect on the probability distribution. Corrections for ties have been thoroughly investigated (see, for example, Noether, 1967, pp. 32–35).

If the ties are handled by the midrank method the variance of W_N in the normal approximation can be corrected to take the ties into account. As we found in (5.7.10), the presence of ties reduces the sum of squares of the ranks by $\sum t(t^2-1)/12$, where t is the number of X and/or Y observations that are tied at any given rank and the sum is over all sets of tied ranks. Substituting this result in (7.3.2) then gives

$$\frac{mn}{N^2(N-1)}\left\{N\left[\frac{N(N+1)(2N+1)}{6}-\frac{\sum t(t^2-1)}{12}\right]-\left[\frac{N(N+1)}{2}\right]^2\right\}$$
$$=\frac{mn(N+1)}{12}-\frac{mn\sum t(t^2-1)}{12N(N-1)} \tag{2.3}$$

The Wilcoxon rank-sum test is actually equivalent to the Mann-Whitney U test discussed in Chapter 6, since a linear relationship exists between the two test statistics. With U defined as the number of times a Y precedes an X, as in (6.6.2), we have

$$U=\sum_{i=1}^{m}\sum_{j=1}^{n}D_{ij}=\sum_{i=1}^{m}(D_{i1}+D_{i2}+\cdots+D_{in})$$

where

$$D_{ij}=\begin{cases}1 & \text{if } Y_j<X_i\\ 0 & \text{if } Y_j>X_i\end{cases}$$

Then $\sum_{j=1}^{n}D_{ij}$ is the number of values of j for which $Y_j<X_i$, or the rank of X_i reduced by n_i, the number of X's which are less than or equal to X_i. Thus we can write

$$\begin{aligned}U=\sum_{i=1}^{m}[r(X_i)-n_i] &= \sum_{i=1}^{m}r(X_i)-(n_1+n_2+\cdots+n_m)\\ &=\sum_{i=1}^{N}iZ_i-(1+2+\cdots+m)\\ &=\sum_{i=1}^{N}iZ_i-\frac{m}{2}(m+1)\\ &=W_N-\frac{m}{2}(m+1)\end{aligned} \tag{2.4}$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

The statistic U (or W_N) can be easily related to the placements introduced in Chapter 2. To see this note that $\sum_{j=1}^{n} D_{ij}$, which counts the total number of Y's that are less than X_i, can be rewritten as $\sum_{i=1}^{n} nG_n(X_i)$, where G_n is the empirical cdf of the Y sample. Now,

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} nG_n(X_i) = \sum_{i=1}^{m}\sum_{j=1}^{n} nG_n(X_{(i)}) = \sum_{i=1}^{m} [\text{rank}(X_{(i)}) - i] \tag{2.5}$$

where rank$(X_{(i)})$ is the rank of the ith-ordered X observation in the combined sample. The last equality in (2.5) also shows that the Mann-Whitney U statistic is a linear function of the Wilcoxon rank-sum test statistic. Thus, all the properties of the two tests are the same, including consistency and the minimum ARE of 0.864 relative to the t test. A confidence-interval procedure based on the Wilcoxon rank-sum test leads to the same results as the one based on the Mann-Whitney test.

The Wilcoxon rank-sum statistic is also equivalent to an ordinary analysis of variance of ranks (see Problem 10.5), a procedure which is easily extended to the case of more than two samples. This situation will be discussed in Chapter 10.

APPLICATIONS

The appropriate rejection regions and P values for the Wilcoxon rank-sum test statistic W_N in terms of $\theta = \mu_Y - \mu_X$ are as follows, where w_O is the observed value of W_N.

Alternative	*Rejection region*	*P value*
$\theta < 0$ $(Y \overset{\text{ST}}{<} X)$	$W_N \geqslant w_\alpha$	$P(W_N \geqslant w_O)$
$\theta > 0$ $(Y \overset{\text{ST}}{>} X)$	$W_N \leqslant w'_\alpha$	$P(W_N \leqslant w_O)$
$\theta \neq 0$	$W_N \geqslant w_{\alpha/2}$ or $W_N \leqslant w'_{\alpha/2}$	2(smaller of above)

The exact cumulative null distribution of W_N is given in Table J for $m \leqslant n \leqslant 10$, as left-tail probabilities for $W_N \leqslant m(N+1)/2$ and right-tail for $W_N \geqslant m(N+1)/2$. For larger sample sizes, the appropriate rejection regions and P values based on the normal approximation with a continuity correction of 0.5 are as follows:

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Alternative	*Rejection region*	*P value*
$\theta < 0$	$W_N \geqslant m(N+1)/2 + 0.5 + z_\alpha\sqrt{mn(N+1)/12}$	$1 - \Phi\left(\frac{w_O - 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12)}}\right)$
$\theta > 0$	$W_N \leqslant m(N+1)/2 - 0.5 - z_\alpha\sqrt{mn(N+1)/12}$	$\Phi\left(\frac{w_O + 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}\right)$
$\theta \neq 0$	Both above with z_α replaced by $z_{\alpha/2}$	2(smaller of above)

If ties are present, the correction for ties derived in (2.3) should be incorporated in the variance term of the rejection regions and P values.

Recall from Section 6.6 that the confidence-interval estimate of $\theta = \mu_Y - \mu_X$ based on the Mann-Whitney test has endpoints which are the $(k+1)$st from the smallest and largest of the mn differences $y_j - x_i$ for all $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$. The value of k is the left-tail rejection region cutoff point of $\alpha/2$ in the null distribution of the Mann-Whitney statistic. Let this $\alpha/2$ cutoff point be $c_{\alpha/2}$. The corresponding cutoff in terms of the Wilcoxon rank-sum statistic from the relationship in (2.3) is $c_{\alpha/2} = w'_{\alpha/2} - m(m+1)/2$. Thus the value of k can be found by subtracting $m(m+1)/2$ from the left-tail critical value of W_N in Table J of the Appendix for the given m and n with $P = \alpha/2$. For example, if $m = 4$, $n = 5$, $P = 0.032$, $w'_{0.032} = 12$, and $c_{0.032} = 12 - 10 = 2$, so that $k + 1 = 3$ and the confidence level is 0.936. Notice that $k + 1$ is always equal to the rank of $w'_{\alpha/2}$ among the entries for the given m and n in Table J because $m(m+1)/2$ is the minimum value of the Wilcoxon rank-sum test statistic W_N.

In practice then, the corresponding confidence interval end points to estimate $\theta = \mu_Y - \mu_X$ are the uth smallest and uth largest of the mn differences $Y_i - X_j$ for all i, j. The appropriate value for u is the rank of that left-tail P among the entries in Table J for the given m and n, for confidence level $1 - 2P$. For m and n outside the range of Table J, we find u from

$$u = \frac{mn}{2} + 0.5 - z_{\alpha/2}\sqrt{\frac{mn(N+1)}{12}} \qquad (2.6)$$

and round down to the next smaller integer if the result is not an integer. Zeros and ties are counted as many times as they occur.

Example 2.1 A time and motion study was made in the permanent mold department at Central Foundry to determine whether there was

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

a pattern to the variation in the time required to pour the molten metal into the die and form a casting of a 6×4 in. Y branch. The metallurgical engineer suspected that pouring times before lunch were shorter than pouring times after lunch on a given day. Twelve independent observations were taken throughout the day, six before lunch and six after lunch. Find the P value for the alternative that mean pouring time before lunch is less than after lunch for the data below on pouring times in seconds.

Before lunch		*After lunch*	
12.6	11.2	16.4	15.4
11.4	9.4	14.1	14.0
13.2	12.0	13.4	11.3

Solution With equal sample sizes $m = n = 6$, either period can be called the X sample. If X denotes pouring time before lunch, the desired alternative is $H_1: \theta = \mu_Y - \mu_X > 0$ and the appropriate P value is in the left tail for W_N. The pooled array with X values underlined is $\underline{9.4}$, $\underline{11.2}$, 11.3, $\underline{11.4}$, $\underline{12.0}$, $\underline{12.6}$, $\underline{13.2}$, 13.4, 14.0, 14.1, 15.4, 16.4, and $W_N = 1 + 2 + 4 + 5 + 6 + 7 = 25$. The P-value is $P(W_N \leqslant 25) = 0.013$ from Table J for $m = 6$, $n = 6$. Thus, the null hypothesis $H_0: \theta = 0$ is rejected in favor of the alternative $H_1: \theta > 0$ at any significance level $\alpha \geqslant 0.013$.

The MINITAB, STATXACT, and SAS solutions to Example 2.1 are shown below. Note that both SAS and MINITAB compute the one-tailed P value as 0.0153 based on the normal approximation with a continuity correction. It is interesting to note that the MINITAB printout includes a confidence interval estimate of $\mu_X - \mu_Y$ that is based on the exact distribution of T_X and this agrees with what we would find (see Problem 8.16), while the test result is based on the normal approximation with a continuity correction. The STATXACT solution gives the exact P value, which agrees with ours, and the asymptotic P value based on the normal approximation without a continuity correction. The SAS solution gives the exact P value, which agrees with ours, and an asymptotic P value based on the normal approximation with a continuity correction (and it tells us so!). SAS also shows a t approximation based on what is called a *rank transformation*. The idea behind a rank transformation is to first replace the original X and Y data values by their ranks in the combined sample and then calculate the usual t statistic from (1.1) using

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

these ranks. The approximate P value is calculated from a t table with $N - 2$ degrees of freedom. The rank transformation idea has been applied to various other classical parametric tests thereby creating new nonparametric tests. The reader is referred to Conover and Iman (1981) for a good introduction to rank transformation and its applications. The SAS output also shows a two-sided result called the Kruskal-Wallis test, which we will cover in Chapter 10.

```
*******************************************
       MINITAB SOLUTION TO EXAMPLE 2.1
*******************************************

Mann-Whitney Test and CI: Before, After

Before     N =   6     Median =      11.700
After      N =   6     Median =      14.050
Point estimate for ETA1-ETA2 is      -2.400
95.5 Percent CI for ETA1-ETA2 is (-4.600,-0.199)
W = 25.0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0153
```

```
*******************************************
       STATXACT SOLUTION TO EXAMPLE 2.1
*******************************************

WILCOXON-MANN-WHITNEY TEST
[  Sum of scores from population <       1 > ]

Summary of Exact distribution of WILCOXON-MANN-WHITNEY statistic:
Min        Max        Mean      Std-dev     Observed  Standardized
21.00      57.00      39.00      6.245       25.00       -2.242
Mann-Whitney Statistic =       4.000

Asymptotic Inference:
One-sided p-value: Pr {   Test Statistic .LE. Observed }  =      0.0125
Two-sided p-value: 2 * One-sided                          =      0.0250

Exact Inference:
One-sided p-value: Pr {   Test Statistic .LE. Observed }  =      0.0130
                   Pr {   Test Statistic .EQ. Observed }  =      0.0054
Two-sided p-value: Pr { | Test Statistic - Mean |
                          .GE.  | Observed - Mean |       =      0.0260
    Two-sided p-value: 2*One-Sided                        =      0.0260
```


MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

```
*******************************
SAS PROGRAM FOR EXAMPLE 2.1
*******************************
DATA TIME;

  INPUT GROUP Time @@;
  DATALINES;
  1 12.6 1 11.2 1 11.4 1 9.4 1 13.2 1 12
  2 16.4 2 15.4 2 14.1 2 14  2 13.4 2 11.3
  ;
  PROC NPAR1WAY DATA=TIME WILCOXON;
  CLASS GROUP;
  VAR TIME;
  EXACT WILLCOXON;
  RUN;
```

```
                  ********************************
                    SAS SOLUTION TO EXAMPLE 2.1
                  ********************************

                     The NPAR1WAY Procedure

               Wilcoxon Scores (Rank Sums) for Variable Time
                   Classified by Variable GROUP

                          Sum of      Expected       Std Dev         Mean
GROUP         N           Scores      Under H0       Under H0        Score
1             6             25.0          39.0       6.244998     4.166667
2             6             53.0          39.0       6.244998     8.833333

                    Wilcoxon Two-Sample Test

               Statistic (S)                 25.0000

                 Normal Approximation
                 Z                            -2.1617
                 One-Sided Pr <  Z             0.0153
                 Two-Sided Pr > |Z|            0.0306

                 t Approximation
                 One-Sided Pr <  Z             0.0268
                 Two-Sided Pr > |Z|            0.0535

                 Exact Test
                 One-Sided Pr <=  S            0.0130
                 Two-Sided Pr >= |S - Mean|    0.0260

                Z includes a continuity correction of 0.5.

                Kruskal-Wallis Test
                Chi-Square           5.0256
                DF                        1
                Pr > Chi-Square      0.0250
```

Example 2.2 In order to compare the relative effectiveness of a calorie-controlled diet and a carbohydrate-controlled diet, eight obese women were divided randomly into two independent groups. Three were placed on a strictly supervised calorie-controlled diet and their total weight losses in 2 weeks were 1, 6, and 7 lb; the others, on a carbohydrate-controlled diet, lost 2, 4, 9, 10, and 12 lb. Find a confidence-interval estimate for the difference in location between Calorie Diet and Carbohydrate Diet, with confidence coefficient near 0.90.

Solution The X sample must be the calorie diet so that $m = 3 \leqslant n = 5$. The example requests a confidence interval for $\mu_X - \mu_Y$. We will proceed by finding a confidence interval on $\mu_Y - \mu_X$ and then take the negative of each endpoint. Table J of the Appendix shows that for $m = 3, n = 5$, the closest we can get to confidence 0.90 is with $P = 0.036$ or exact confidence level 0.928; this entry has rank 2 so that $u = 2$. The $3(5) = 15$ differences $Y - X$ are shown below. The second smallest difference is -4 and the second largest difference is 9, or $-4 \leqslant \mu_Y - \mu_X \leqslant 9$; the corresponding negative interval is $-9 \leqslant \mu_X - \mu_Y \leqslant 4$. Notice that by listing the Y values in an array and then subtracting successively larger X values, the smallest and largest differences are easy to find.

Y	$Y-1$	$Y-6$	$Y-7$
2	1	−4	−5
4	3	−2	−3
9	8	3	2
10	9	4	3
12	11	6	5

The MINITAB and STATXACT solutions to this example are shown below. Note that the MINITAB output shows a confidence level 0.926, which is almost equal to the exact level, 0.928. The MINITAB confidence limits also do not match exactly with ours but are very close. The STATXACT solution matches exactly with ours, although the output does not seem to indicate what the exact confidence level is. It is interesting to note that for this example, with such small sample sizes, the exact and the asymptotic methods produced identical results. Note also that STATXACT calls this procedure the Hodges-Lehmann estimate of the shift parameter.

```
            *******************************************
              MINITAB SOLUTION TO EXAMPLE 2.2
            *******************************************

MTB > Mann-Whitney 90.0 C1 C2;

SUBC>   Alternative 0.

Mann-Whitney Test and CI: C1, C2

C1         N =   3    Median =       6.000
C2         N =   5    Median =       9.000
Point estimate for ETA1-ETA2 is     -3.000

92.6 Percent CI for ETA1-ETA2 is (-8.998,3.997)
W = 10.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.3711

Cannot reject at alpha = 0.05

            *******************************************
                STATXACT SOLUTION TO EXAMPLE 2.2
            *******************************************

HODGES-LEHMANN ESTIMATES OF SHIFT PARAMETER

POP_1 :        1             POP_2 :        2

Summary of WILCOXON MANN-WHITNEY statistic for POP_1

Min          Max          Mean       Std-dev     Observed   Standardized

6.000        21.00        13.50        3.354        10.00        -1.043
Mann-Whitney Statistic =         4.000
Point Estimate of Shift : Theta = POP_1 - POP_2 =        -3.000

90.00% Confidence Interval for Theta :
          Asymptotic : (      -9.000 ,       4.000)
          Exact      : (      -9.000 ,       4.000)
```

8.3 OTHER LOCATION TESTS

Generally, almost any set of monotone-increasing weights a_i, which are adopted for a linear rank statistic, will provide a consistent test for shift in location. Only a few of the better-known ones will be covered here.

TERRY-HOEFFDING (NORMAL SCORES) TEST

The *Terry* (1952) and *Hoeffding* (1951) or the *Fisher-Yates normal scores test* uses the weights $a_i = E(\xi_{(i)})$, where $\xi_{(i)}$ is the ith-order

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

statistic from a standard normal population; the linear rank test statistic is

$$c_1 = \sum_{i=1}^{N} E(\xi_{(i)}) Z_i \tag{3.1}$$

These expected values of standard normal order statistics are tabulated for $N \leqslant 100$ and some larger sample sizes in Harter (1961), so that the exact null distribution can be found by enumeration. Tables of the distribution of the test statistic are given in Terry (1952) and Klotz (1964). The Terry test statistic is symmetric about the origin, and its variance is

$$\sigma^2 = mn \frac{\sum_{i=1}^{N} [E(\xi_{(i)})]^2}{N(N-1)} \tag{3.2}$$

The normal distribution provides a good approximation to the null distribution for larger sample sizes. An approximation based on the t distribution is even closer. This statistic is $t = r(N-2)^{1/2}/(1-r^2)^{1/2}$, where $r = c_1/[\sigma^2(N-1)]^{1/2}$ and the distribution is approximately Student's t with $N-2$ degrees of freedom.

The Terry test is asymptotically optimum against the alternative that the populations are both normal distributions with the same variance but different means. Under the classical assumptions for a test of location then, its ARE is 1 relative to Student's t test. For certain other families of continuous distributions, the Terry test is more efficient than Student's t test (ARE > 1) (Chernoff and Savage, 1958).

The weights employed for the Terry test $E(\xi_{(i)})$ are often called *expected normal scores*, since the order statistics of a sample from the standard normal population are commonly referred to as normal scores. The idea of using expected normal scores instead of integer ranks as rank-order statistics is appealing generally, since for many populations the expected normal scores may be more "representative" of the raw data or variate values. This could be investigated by comparing the correlation coefficients between (1) variate values and expected normal scores and (2) variate values and integer ranks for particular families of distributions. The limiting value of the correlation between variate values from a normal population and the expected normal scores is equal to 1, for example (see Section 5.5). Since inferences based on rank-order statistics are really conclusions about transformed variate values, that transformation which most closely approximates the actual data should be most efficient when these inferences are extended to the actual data.

Since the Terry test statistic is the sum of the expected normal scores of the variables in the X sample, it may be interpreted as identical to the Wilcoxon rank-sum test of the previous section when the normal-scores transformation is used instead of the integer-rank transformation. Other linear rank statistics for location can be formed in the same way by using different sets of rank-order statistics for the combined samples. An obvious possibility suggested by the Terry test is to use the scores $\Phi^{-1}[i/(N+1)]$, where $\Phi(x)$ is the cumulative standard normal distribution, since we showed in Chapter 2 that $\Phi^{-1}[i/(N+1)]$ is a first approximation to $E(\xi_{(i)})$. If κ_p is the pth quantile point of the standard normal distribution, $\Phi(\kappa_p) = p$ and $\kappa_p = \Phi^{-1}(p)$. Therefore here the ith-order statistic in the combined ordered sample is replaced by the $[i/(N+1)]$st quantile point of the standard normal. This is usually called the *inverse-normal-scores transformation* and forms the basis of the following test.

VAN DER WAERDEN TEST

When the inverse normal scores are used in forming a linear rank statistic, we obtain the *van der Waerden* (1952, 1953) X_N *test*, where

$$X_N = \sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right) Z_i \tag{3.3}$$

In other words, the constant a_i in a general linear rank statistic is given by the value on the abscissa of the graph of a standard normal density function such that the area to the left of a_i is equal to $i/(N+1)$. These weights a_i are easily found from tables of the cumulative normal distribution. Tables of critical values are given in van der Waerden and Nievergelt (1956) for $N \leqslant 50$. The X_N statistic is symmetric about zero and has variance

$$\sigma^2 = mn \frac{\sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2}{N(N-1)} \tag{3.4}$$

For larger sample sizes, the null distribution of the standardized X_N is well approximated by the standard normal distribution.

The X_N test is perhaps easier to use than the Terry test because the weights are easily found for any N. Otherwise, there is little basis

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

for a choice between them. In fact, the van der Waerden test is asymptotically equivalent to the Terry test. Since $\mathrm{var}(\xi_{(i)}) \to 0$ as $N \to \infty$, $\xi_{(i)}$ converges in probability to $E(\xi_{(i)})$, the weights for the Terry test statistic. However, by the probability-integral transformation, $\Phi(\xi_{(i)})$ is the ith-order statistic of a sample of size N from the uniform distribution. Therefore from (2.8.2) and (2.8.3), $E[\Phi(\xi_{(i)})] = i/(N+1)$ and

$$\mathrm{var}[\Phi(\xi_{(i)})] = \frac{i(N-i+1)}{(N+1)^2(N+2)} \to 0$$

as $N \to \infty$. This implies that $\Phi(\xi_{(i)})$ converges in probability to $i/(N+1)$ and $\xi_{(i)}$ converges to $\Phi^{-1}[i/(N+1)]$. We may conclude that the expected normal scores and the corresponding inverse normal scores are identical for all N as $N \to \infty$. Thus, the large sample properties, including the ARE, are the same for the Terry and the van der Waerden tests.

It should be noted that the expected normal scores and inverse normal scores may be useful in any procedures based on rank-order statistics. For example, in the one-sample and paired-sample Wilcoxon signed-rank test discussed in Section 5.7, the rank of the absolute value of the difference $|D_i|$ can be replaced by the corresponding expected value of the absolute value of the normal score $E(|\xi_{(i)}|)$ (which is not equal to the absolute value of the expected normal score). The sum of those "ranks" which correspond to positive differences D_i is then employed as a test statistic. This statistic provides the asymptotically optimum test of location when the population of differences is normal and thus has an ARE of 1 relative to Student's t test in this case. Expected normal scores are also useful in rank-correlation methods, which will be covered in Chapter 11.

Example 2.3 We illustrate the Terry and van der Waerden tests using data from Example 2.1 on pouring times with $m = 6$ and $n = 6$. The first six expected normal scores for the Terry test with $\mathrm{N} = 12$ are $-1.6292, -1.1157, -0.7928, -0.5368, -0.3122$, and -0.1026; the other six are the same values but with positive signs by symmetry. For example, the seventh expected normal score is 0.1026, the eighth is 0.3122, and so on. We calculate $c_1 = -3.5939$ from (3.1) with variance $\sigma^2 = 2.6857$ from (3.2). The z statistic for the normal approximation is $z = -2.1930$ with a one-sided P value $P(Z \leqslant -2.1930) = 0.0142$. For the van der Waerden test, the first six inverse normal scores with

$N = 12$ are $-1.4261, -1.0201, -0.7363, -0.5024, -0.2934$, and -0.0966; the remaining six are the same values but with positive signs by symmetry. The test statistic is $X_N = -3.242$ from (3.3), with variance $\sigma^2 = 2.1624$, from (3.4). The z statistic for the normal approximation is $z = -2.2047$ with a one-tailed P value $P(Z \leqslant -2.2047) = 0.0137$.

Note that we do not use a continuity correction in calculating either of these two z statistics, because the weights a_i for both of these test statistics are continuous variables and not integers.

The SAS solution for the data in Example 2.1 using the van der Waerden test is shown below. Note that it agrees exactly with ours. STATXACT has an option called the normal scores test, but it uses the inverse normal scores as weights, as opposed to the expected normal scores. In other words, it calculates the van der Waerden statistic. This solution is also shown below. Note also that both SAS and STATXACT also provide exact P-values corresponding to the test statistic -3.242 and this one-tailed P-value is identical to the one found earlier in Example 1.1 for the Wilcoxon rank-sum statistic.

****
SAS SOLUTION TO EXAMPLE 2.1
****

Program:

```
DATA TIME;
INPUT GROUP TIME @@;
DATALINES;
1 12.6 1 11.2 1 11.4 1 9.4 1 13.2 1 12
2 16.4 2 15.4 2 14.1 2 14 2 13.4 2 11.3
;
PROC NPAR1WAY DATA=TIME VW;
CLASS GROUP;
VAR TIME;
EXACT VW;
RUN;
```

Output

The NPAR1WAY Procedure

Van der Waerden Scores (Normal) for Variable Time
Classified by Variable GROUP

GROUP	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	6	-3.241937	0.0	1.470476	-0.540323
2	6	3.241937	0.0	1.470476	0.540323

```
Van der Waerden Two-Sample Test

Statistic (S)                   -3.2419

Normal Approximation
Z                               -2.2047
One-Sided Pr <  Z                0.0137
Two-Sided Pr > |Z|               0.0275

Exact Test
One-Sided Pr <=  S               0.0130
Two-Sided Pr >= |S - Mean|       0.0260

Van der Waerden One-Way Analysis

Chi-Square          4.8606
DF                       1
Pr > Chi-Square     0.0275
```

```
*******************************
STATXACT SOLUTION TO EXAMPLE 2.1
*******************************
```

```
NORMAL SCORES TEST

[  Sum of scores from population <       1 > ]

Summary of Exact distribution of NORMAL SCORES statistic:

Min          Max          Mean       Std-dev     Observed   Standardized
-4.075     4.075  -7.772e-016        1.470       -3.242         -2.205

Asymptotic Inference:
One-sided p-value: Pr {   Test Statistic .LE. Observed }  =     0.0137
Two-sided p-value: 2 * One-sided                          =     0.0275

Exact Inference:
One-sided p-value: Pr {   Test Statistic .LE. Observed }  =     0.0130
                   Pr {   Test Statistic .EQ. Observed }  =     0.0011
Two-sided p-value: Pr { | Test Statistic - Mean |
                            .GE.  | Observed - Mean |     =     0.0260
Two-sided p-value: 2*One-Sided                            =     0.0260
```

PERCENTILE MODIFIED RANK TESTS

Another interesting linear rank statistic for the two-sample location problem is a member of the class of so-called *percentile modified linear rank tests* (Gastwirth, 1965). The idea is as follows. We select two numbers s and r, both between 0 and 1, and then score only the data in the upper sth and lower rth percentiles of the combined sample.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

In other words, a linear rank statistic is formed in the usual way except that a score of zero is assigned to a group of observations in the middle of the combined array. Symbolically, we let $S = [Ns] + 1$ and $R = [Nr] + 1$, where $[x]$ denotes the largest integer not exceeding the number x. Define B_r and T_s, as

N odd:

$$B_r = \sum_{i=1}^{R} (R - i + 1) Z_i$$

and

$$T_s = \sum_{i=N-S+1}^{N} [i - (N - S)] Z_i$$

N even: (3.3)

$$B_r = \sum_{i=1}^{R} (R - i + 1/2) Z_i$$

and

$$T_s = \sum_{i=N-S+1}^{N} [i - (N - S) - 1/2] Z_i$$

The combination $T_s - B_r$ provides a test for location, and $T_s + B_r$ is a test for scale, which will be discussed in the next chapter. It is easily seen that if N is even and $S = R = N/2$, so that no observations are assigned a score of zero, $T_s - B_r$ is equivalent to the Wilcoxon test. When N is odd and all the sample data are used, the tests differ slightly because of the different way of handing the middle observation $z_{(N+1)/2}$.

The mean and variance of the $T_s \pm B_r$ statistics can be calculating using Theorem 7.3.2 alone if $S + R \leqslant N$, remembering that $a_i = 0$ for $R + 1 \leqslant i \leqslant N - S$. Alternatively, Theorems 7.3.2 and 7.3.3 can be used on the pieces T_s and B_r along with the fact that

$$\text{var}(T_s \pm B_r) = \text{var}(T_s) + \text{var}(B_r) \pm 2\text{cov}(T_s, B_r)$$

The results for N even and $S = R$ are

$$E(T_s - B_r) = 0 \qquad \text{var}(T_s - B_r) = \frac{mnS(4S^2 - 1)}{6N(N - 1)} \tag{3.4}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

By Theorem 7.3.4, the null distribution of $T_s - B_r$ is symmetric about the origin for any m and n when $S = R$. Tables of the null distribution for $m = n \leqslant 6$ are given in Gibbons and Gastwirth (1966). It is also shown there empirically that for significance levels not too small, say at least 0.025, the normal distribution may be used to define critical regions with sufficient accuracy for most practical purposes when $m = n \geqslant 6$.

One of the main advantages of this test is that a judicious choice of s and r may lead to a test which attains higher power than the Wilcoxon rank-sum test without having to introduce complicated scoring systems. For example, any knowledge of asymmetry in the populations might be incorporated into the test statistic. The asymptotic relative efficiency of this test against normal alternatives reaches its maximum value of 0.968 when $s = r = 0.42$; when $s = r = 0.5$, the ARE is 0.955, as for the Wilcoxon rank-sum statistic.

8.4 SUMMARY

In this chapter we covered several additional tests for the two-sample problem; all of them are linear rank tests. The two-sample tests covered in Chapter 6 are appropriate for general alternatives that do not specify any particular kind of difference between the population distributions. The tests covered in this chapter are especially appropriate for the location alternative.

The Wilcoxon rank-sum test of Section 8.2 is by far the best known two-sample nonparametric test for location, and it is equivalent to the Mann-Whitney U test covered in Section 6.6. The discussion of power given there applies equally here. The expressions given there for sample size determination also apply here. Other tests for location covered in this chapter are the Terry-Hoeffding expected normal scores test and the van der Waerden inverse normal scores test. These two tests are asymptotically equivalent and their asymptotic relative efficiency is one relative to the normal theory test for normal distributions. Thus their ARE is somewhat higher than that of the Wilcoxon test for normal distributions, but they can have lower power for other distributions. These other tests are not as convenient to use as the Wilcoxon test and are not very well known. Further, they do not have a convenient procedure for finding a corresponding confidence interval estimate for the difference in the medians. Finally, we cover the percentile modified rank tests for location, which are simply generalizations of the Wilcoxon rank-sum test.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

PROBLEMS

8.1. Given independent samples of m X and n Y variables, define the following random variables for $i = 1, 2, \ldots, m$:

$$K_i = \text{rank of } X_i \text{ among } X_1, X_2, \ldots, X_m$$

$$R_i = \text{rank of } X_i \text{ among } X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$$

Use K_i and R_i to prove the linear relationship between the Mann-Whitney and Wilcoxon rank-sum statistics given in (2.4).

8.2. A single random sample $D_1, D_2, \ldots, D_N$ of size N is drawn from a population which is continuous and symmetric. Assume there are m positive values, n negative values, and no zero values. Define the $m + n = N$ random variables

$$X_i = D_i \qquad \text{if } D_i > 0$$
$$Y_i = |D_i| \qquad \text{if } D_i < 0$$

Then the $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ constitute two independent random samples of sizes m and n.

(a) Show that the two-sample Wilcoxon rank-sum statistic W_N of (2.1) for these two samples equals the Wilcoxon signed-rank statistic T^+ defined in (5.7.1).

(b) If these two samples are from identical populations, the median of the symmetric D population must be zero. Therefore the null distribution of W_N is identical to the null distribution of T^+ conditional upon the observed number of plus and minus signs. Explain fully how tables of the null distribution of W_N could be used to find the null distribution of T^+. Since for N large, m and n will both converge to the constant value $N/2$ in the null case, these two test statistics have equivalent properties asymptotically.

8.3. Generate by enumeration the exact null probability distribution of $T_s - B_r$ as defined in (3.3) for $m = n = 3$, all $S = R < 3$, and compare the rejection regions for $\alpha \leqslant 0.10$ with those for the Wilcoxon rank-sum test W_N when $m = n = 3$.

8.4. Verify the results given in (3.4) for the mean and variance of $T_s - B_r$ when $S = R$ and N is even and derive a similar result for $S = R$ when N is odd.

8.5. Show that the median test of Section 6.4 is a linear rank test.

8.6. Giambra and Quilter (1989) performed a study to investigate gender and age difference in ability to sustain attention when given Mackworth's Clock-Test. This clock is metal with a plain white face and a black pointer that moves around the face in 100 discrete steps of 36 degrees each. During the test period the pointer made 23 double jumps, defined as moving twice the normal distance or 7.2 degrees in the same time period, at random and irregular intervals. Subjects were told that double jumps would occur and asked to signal their recognition of occurrence by pressing a button. Scores were the number of correct recognitions of the double jumps. The scores below are for 10 men in age groups 18–29 and 10 men in age group 50–59. Determine whether median number of correct scores is larger for young men than for older men.

Age $18 - 29$: $11, 13, 15, 15, 17, 19, 20, 21, 21, 22$
Age $50 - 59$: $8, 9, 10, 11, 12, 13, 5, 17, 19, 23$

8.7. Elam (1988) conducted a double-blind study of 18 adult males to investigate the effects of physical resistance exercises and amino acid dietary supplements on body

Table 1

Treatment group				Control group			
Subject	*Mass*	*Fat*	*Girth*	*Subject*	*Mass*	*Fat*	*Girth*
1	−2.00	1.14	−17.00	1	1.00	−0.56	11.00
2	0.00	−2.64	2.00	2	0.50	0.87	5.00
3	−1.00	−1.96	23.00	3	−0.75	−0.75	1.00
4	−4.00	0.86	13.00	4	−2.00	−0.60	35.00
5	−0.75	−2.35	2.00	5	−3.00	0.00	−5.00
6	−1.75	−2.51	5.00	6	−2.50	−2.54	2.00
7	−2.75	0.55	8.00	7	0.00	−3.10	3.00
8	0.00	3.40	3.00	8	0.25	3.48	−7.00
9	−1.75	0.00	7.00				
10	1.00	−4.94	10.00				

mass, body fat, and composite girth. Ten of the subjects received the diet supplement and eight received a placebo. All subjects participated in 15 resistance exercise workouts of one hour each spread over a 5-week period. Workloads were tailored to abilities of the individual subjects but escalated in intensity over the period. The data in Table 1 are the changes (after minus before) in body mass, body fat, and composite body girth for the amino acid (Treatment) group and placebo (Control) group of subjects. Were amino acid supplements effective in reducing the body mass (kg), fat (%), and girth (cm)?

8.8. Howard, Murphy, and Thomas (1986) (see Problem 5.12) also investigated whether pretest anxiety scores differed for students enrolled in two different sections of the introduction to computer courses. Seven students were enrolled in each section, and the data are shown below. Is there a difference in median scores?

Section 1: 20, 32, 22, 21, 27, 26, 38
Section 2: 34, 20, 30, 28, 25, 23, 29

8.9. A travel agency wanted to compare the noncorporate prices charged by two major motel chains for a standard-quality single room at major airport locations around the country. A random sample of five Best Eastern motels and an independent sample of six Travelers' Inn motels, all located at major airports, gave the approximate current total costs of a standard single room as shown below. Find a 95% confidence interval estimate of the difference between median costs at Best Eastern and Travelers' Inn motels.

Best Eastern: $68, 75, 92, 79, 95
Travelers' Inn: $69, 76, 81, 72, 75, 80

8.10. Smokers are commonly thought of as nervous people whose emotionality is at least partly caused by smoking because of the stimulating effect tobacco has on the nervous system. Nesbitt (1972) conducted a study with 300 college students and concluded that smokers are less emotional than nonsmokers, that smokers are better able to tolerate the physiological effects of anxiety, and that, over time, smokers become less emotional than nonsmokers. Subjects of both sexes were drawn from three different colleges and classified as smokers if they smoked any number of cigarettes on a regular basis. In one aspect of the experiment all subjects were given the Activity Preference Questionnaire (APQ), a test designed to measure the emotionality of the subjects. The APQ is scored using an ordinal scale of 0–33, with lower scores indicating less emotionality, that is, greater sociopathy. The mean overall scores were 18.0 for smokers and 20.3 for nonsmokers. Suppose this experiment is repeated using a group of only 8 randomly chosen smokers and 10 randomly chosen nonsmokers, and the score results are shown below. Do these data support the same conclusion concerning emotionality as Dr. Nesbitt's data?

Smokers:	16, 18, 21, 14, 25, 24, 27, 12
Nonsmokers:	17, 15, 28, 31, 30, 26, 27, 20, 21, 19

8.11. A group of 20 mice are allocated to individual cages randomly. The cages are assigned in equal numbers, randomly, to two treatments, a control A and a certain drug B. All animals are infected, in a random sequence, with tuberculosis. The number of days until the mice die after infection are given as follows (one mouse got lost):

Control A:	5, 6, 7, 7, 8, 8, 9, 12
Drug B:	7, 8, 8, 8, 9, 9, 12, 13, 14, 17

Since a preliminary experiment has established that the drug is not toxic, we can assume that the drug group cannot be worse (die sooner) than the control group under any reasonable conditions. Test the null hypothesis that the drug is without effect at a significance level of 0.05 and briefly justify your choice of test.

8.12. The following data represent two independent random samples drawn from continuous populations which are thought to have the same form but possibly different locations.

X:	79, 13, 138, 129, 59, 76, 75, 53
Y:	96, 141, 133, 107, 102, 129, 110, 104

Using a significance level not exceeding 0.10, test

(a) The null hypothesis that the two populations are identical and find the P value (Do not use an approximate test)

(b) The null hypothesis that the locations are the same and find the appropriate one-tailed P value

8.13. A problem of considerable import to the small-scale farmer who purchases young pigs to fatten and sell for slaughter is whether there is any difference in weight gain for male and female pigs when the two genders are subjected to identical feeding treatments. If there is a difference, the farmer can optimize production by buying only one gender of pigs for fattening. As a public service, an agricultural experiment station decides to run a controlled experiment to determine whether gender is an important factor in weight gain. They placed 8 young male pigs in one pen 8 young females in another pen and gave each pen identical feeding treatments for a fixed period of time. The initial weights were all between 35 and 50 lb, and the amounts of weight gain in pounds for the two genders are recorded below. Unfortunately, one of the female pigs died so there are only 7 observations in that group. Analyze the data below using both a test and a confidence-interval approach with confidence coefficient near 0.90.

Female pigs:	9.31, 9.57, 10.21, 8.86, 8.52, 10.53, 9.21
Male pigs:	9.14, 9.98, 8.46, 8.93, 10.14, 10.17, 11.04, 9.43

8.14. How would you find the confidence-interval end points for the parameter of interest when the interval has confidence level nearest 0.90 and corresponds to:

(a) The sign text with $n = 11$
(b) The Wilcoxon signed-rank test with $n = 11$
(c) The Wilcoxon rank-sum test with $m = 5, n = 6$

In each case define the function Z of the observations, give the numerical values of L and U for the order statistics $Z(L)$ and $Z(U)$, and give the exact confidence level.

8.15. A self-concept test was given to a random sample consisting of six normal subjects and three subjects under psychiatric care. Higher scores indicate more self-esteem. The data are as follows:

Normal:	$62, 68, 78, 92, 53, 81$
Psychiatric:	$54, 70, 79$

(a) Find a P value relevant to the alternative that psychiatric patients have lower self-esteem than normal patients.

(b) Find a confidence interval for the difference of the locations (level nearest 0.90).

8.16. Verify the confidence interval estimate of $\mu_X - \mu_Y$ with exact confidence coefficient at least 0.95 given in the MINITAB solution to Example 2.1.

9

Linear Rank Tests for the Scale Problem

9.1 INTRODUCTION

Consider again the situation of Chapter 8, where the null hypothesis is that two independent samples are drawn from identical populations; however, now suppose that we are interested in detecting differences in variability or dispersion instead of location. Some of the tests presented in Chapters 6 and 8, namely, the median, Mann-Whitney, Wilcoxon rank-sum, Terry, van der Waerden, and $T_S - B_r$ tests, were noted to be particularly sensitive to differences in location when the populations are identical otherwise, a situation described by the relation $F_Y(x) = F_X(x-\theta)$. These tests cannot be expected to perform especially well against other alternatives. The general two-sample tests of Chapter 6, like the Wald-Wolfowitz runs test or Kolmogorov-Smirnov tests, are affected by any type of difference in the populations and therefore cannot be relied upon as efficient for detecting differences in variability. Some other nonparametric tests are needed for the dispersion problem.

The classical test for which we are seeking an analog is the test for equality of variances, H_0: $\sigma_X = \sigma_Y$, against one- or two-sided alternatives. If it is reasonable to assume that the two populations are both normal distributions, the parametric test statistic is

$$F_{m-1,n-1} = \sum_{i=1}^{m} \frac{(X_i - \bar{X})^2}{m-1} \Big/ \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}$$

which has Snedecor's F distribution with $m-1$ and $n-1$ degrees of freedom. The F test is not particularly robust with respect to the normality assumption. If there is reason to question the assumptions inherent in the construction of the test, a nonparametric test of dispersion is appropriate.

The F test does not require any assumption regarding the locations of the two normal populations. The magnitudes of the two sample variances are directly comparable since they are each computed as measures of deviations around the respective sample means. The traditional concept of dispersion is a measure of spread around some population central value. The model for the relationship between the two normal populations assumed for the F test might be written

$$F_{Y-\mu_Y}(x) = F_{X-\mu_X}\left(\frac{\sigma_X}{\sigma_Y}x\right) = F_{X-\mu_X}(\theta x) \qquad \text{for all } x \text{ and some } \theta > 0 \tag{1.1}$$

where $\theta = \sigma_X/\sigma_Y$ and $F_{(X-\mu_X)/\sigma_X}(x) = \Phi(x)$, and the null hypothesis to be tested is H_0: $\theta = 1$. We could say then that we assume that the distributions of $X - \mu_X$ and $Y - \mu_Y$ differ only by the scale factor θ for any μ_X and μ_Y, which need not be specified. The relationship between the respective moments is

$$E(X - \mu_X) = \theta E(Y - \mu_Y) \qquad \text{and} \qquad \text{var}(X) = \theta^2 \ \text{var}(Y)$$

Since medians are the customary location parameters in distribution-free procedures, if nonparametric dispersion is defined as spread around the respective medians, the nonparametric model corresponding to (1.1) is

$$F_{Y-M_Y}(x) = F_{X-M_X}(\theta x) \qquad \text{for all } x \text{ and some } \theta > 0 \tag{1.2}$$

Suppose that the test criterion we wish to formulate for this model is to be based on the configuration of the X and Y random variables in the combined ordered sample, as in a linear rank test. The characteristics of respective locations and dispersions are inextricably

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

mixed in the combined sample ordering, and possible location differences may mask dispersion differences. If the population medians are known, the model (1.2) suggests that the sample observations should be adjusted by

$$X_i' = X_i - M_X \quad \text{and} \quad Y_j' = Y_j - M_Y \qquad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n$$

Then the X_i' and Y_j' populations both have zero medians, and the arrangement of X' and Y' random variables in the combined ordered sample should indicate dispersion differences as unaffected by location differences. The model is then $F_Y(x) = F_X(\theta x)$. In practice, M_X and M_Y would probably not be known, so that this is not a workable approach. If we simply assume that $M_X = M_Y = M$ unspecified, the combined sample arrangement of the unadjusted X and Y should still reflect dispersion differences. Since the X and Y populations differ only in scale, the logical model for this situation would seem to be the alternative

$$H_S\colon F_Y(x) = F_X(\theta x) \qquad \text{for all } x \text{ and some } \theta > 0,\ \theta \neq 1 \tag{1.3}$$

This is appropriately called the *scale alternative* because the cumulative distribution function of the Y population is the same as that of the X population but with a compressed or enlarged scale according as $\theta > 1$ or $\theta < 1$, respectively.

In Figure 1.1a, the relation $F_Y(x) = F_X(\theta x)$ is shown for $F_Y(x) = \Phi(x)$, the standard normal, and $\theta > 1$. Since $\mu_X = M_X = \mu_Y = 0$ and $\theta = \sigma_X/\sigma_Y$, this model is a special case of (1.1) and (1.2).

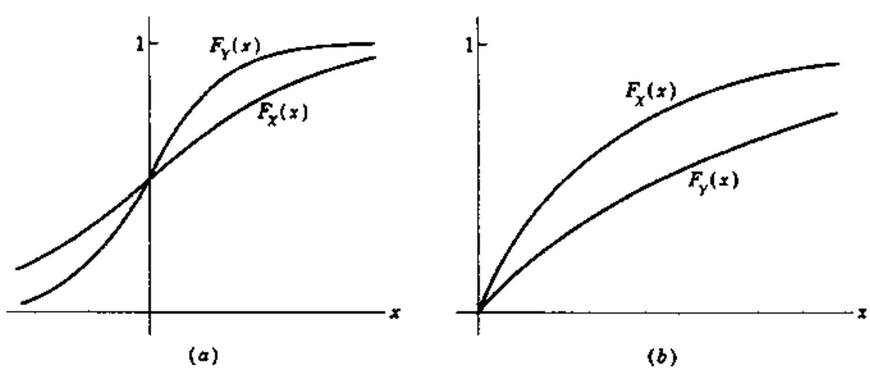


Fig. 1.1 $F_Y(x) = F_X(\theta x)$. (*a*) F_X normal, $\theta > 1$; (*b*) F_X exponential, $\theta < 1$.

Figure 1.1b illustrates the difficulty in thinking any arbitrary distribution may be taken for the scale alternative in (1.3) to be interpreted exclusively as a dispersion alternative. Here we have a representation of the exponential distribution in H_S for $\theta < 1$, for example, $f_X(x) = e^{-x}, x > 0$, so that $f_Y(x) = \theta\, e^{-\theta x}$ for some $\theta < 1$. Since $\mathrm{var}(X) = 1$ and $\mathrm{var}(Y) = 1/\theta^2$, it is true that $\sigma_Y > \sigma_X$. However, $E(X) = 1$ and $E(Y) = 1/\theta > E(X)$, and further $M_X = \ln 2$ while $M_Y = \ln(2/\theta) > M_X$ for all $\theta < 1$. The combined ordered arrangement of samples from these exponential populations will be reflective of both the location and dispersion differences. The scale alternative in (1.3) should be interpreted as a dispersion alternative only if the population locations are zero or very close to zero.

Actually, the scale model $F_Y(x) = F_X(\theta x)$ is not general enough even when the locations are the same. This relationship implies that $E(X) = \theta E(Y)$ and $M_X = \theta M_Y$, so that the locations are identical for all θ only if $\mu_X = \mu_Y = 0$ or $M_X = M_Y = 0$. A more general scale alternative can be written in the form

$$H_S\colon F_{Y-M}(x) = F_{X-M}(\theta x) \qquad \text{for all } x \text{ and some } \theta > 0, \theta \neq 1 \tag{1.4}$$

where M is interpreted to be the common median. Both (1.3) and (1.4) are called the scale alternatives applicable to the two-sample scale problem, but in (1.3) we essentially assume without loss of generality that $M = 0$.

Many tests based on the ranks of the observations in a combined ordering of the two samples have been proposed for the scale problem. If they are to be useful for detecting differences in dispersion, we must assume either that the medians (or means) of the two populations are equal but unknown or that the sample observations can be adjusted to have equal locations, by subtracting the respective location parameters from one set. Under these assumptions, an appropriate set of weights for a linear rank-test statistic will provide information about the relative spread of the observations about their common central value. If the X population has a larger dispersion, the X values should be positioned approximately symmetrically at both extremes of the Y values. Therefore the weights a_i should be symmetric, for example, small weights in the middle and large at the two extremes, or vice versa. We shall consider several choices for simple sets of weights of this type which provide linear rank tests particularly sensitive to scale differences only. These are basically the best-known tests—the Mood test, the Freund-Ansari-Bradley-David-Barton tests, the Siegel-Tukey test, the Klotz normal-scores test, the percentile modified rank tests, and the Sukhatme test. Many other tests have also been proposed in the literature; some of these are covered in Section 9.9. Duran (1976) gives a survey of nonparametric scale tests. Procedures for finding confidence interval estimates of relative scale are covered in Section 9.8. Examples and applications are given in Section 9.10.

9.2 THE MOOD TEST

In the combined ordered sample of N variables with no ties, the average rank is the mean of the first N integers, $(N+1)/2$. The deviation of the rank of the ith ordered variable about its mean rank is $i-(N+1)/2$, and the amount of deviation is an indication of relative spread. However, as in the case of defining a measure of sample dispersion in classical descriptive statistics, the fact that the deviations are equally divided between positive and negative numbers presents a problem in using these actual deviations as weights in constructing a linear rank statistics. For example, if Z_i is the usual indicator variable for the X observations and $m=n=3$, the ordered arrangements

$$XYXYXY \qquad \text{and} \qquad XXYYYX$$

both have $\sum_{i=1}^{6}\left(i-\frac{N+1}{2}\right)Z_i=-1.5$, but the first arrangement suggests the variances are equal and the second suggests the X's are more dispersed than the Y's. The natural solution is to use as weights either the absolute values or the squared values of the deviations to give equal weight to deviations on either side of the central value.

The *Mood* (1954) *test* is based on the sum of squares of the deviations of the X ranks from the average combined rank, or

$$M_N=\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 Z_i \tag{2.1}$$

A large value of M_N would imply that the X's are more widely dispersed, and M_N small implies the opposite conclusion. Specifically, the set of weights is as shown in Tables 2.1 and 2.2 for N even and N odd, respectively. The larger weights are in the tails of the arrangement. When N is odd, the median of the combined sample is assigned a weight of zero. In that case, therefore, the middle observation is essentially ignored, but this is necessary to achieve perfectly symmetric weights.

Table 2.1 Mood test weights for N even

i	1	2	3	$\dots$	$\frac{N}{2}-1$	$\frac{N}{2}$
a_i	$\left(\frac{N-1}{2}\right)^2$	$\left(\frac{N-3}{2}\right)^2$	$\left(\frac{N-5}{2}\right)^2$	$\dots$	$\left(\frac{3}{2}\right)^2$	$\left(\frac{1}{2}\right)^2$
i	$\frac{N}{2}+1$	$\frac{N}{2}+2$	$\dots$	$N-2$	$N-1$	N
a_i	$\left(\frac{1}{2}\right)^2$	$\left(\frac{3}{2}\right)^2$	$\dots$	$\left(\frac{N-5}{2}\right)^2$	$\left(\frac{N-3}{2}\right)^2$	$\left(\frac{N-1}{2}\right)^2$

Table 2.2 Mood test weights for N odd

i	1	2	3	$\dots$	$\frac{N-1}{2}$	$\frac{N+1}{2}$
a_i	$\left(\frac{N-1}{2}\right)^2$	$\left(\frac{N-3}{2}\right)^2$	$\left(\frac{N-5}{2}\right)^2$	$\dots$	$(1)^2$	0
i	$\frac{N+3}{2}$	$\dots$	$N-2$	$N-1$	N	
a_i	$(1)^2$	$\dots$	$\left(\frac{N-5}{2}\right)^2$	$\left(\frac{N-3}{2}\right)^2$	$\left(\frac{N-1}{2}\right)^2$	

The moments of M_N under the null hypothesis are easily found from Theorem 7.3.2 (Section 7.3) as follows:

$$\begin{aligned}
NE(M_N) &= m\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 \\
&= m\left[\sum i^2-(N+1)\sum i+\frac{N(N+1)^2}{4}\right] \\
&= m\left[\frac{N(N+1)(2N+1)}{6}-\frac{N(N+1)^2}{2}+\frac{N(N+1)^2}{4}\right]
\end{aligned}$$

Then $12NE(M_N)=mN(N+1)(N-1)$ and

$$E(M_N)=\frac{m(N^2-1)}{12} \tag{2.2}$$

Further,

$$\begin{aligned}
&N^2(N-1)\mathrm{var}(M_N) \\
&\quad = mn\left\{N\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^4-\left[\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2\right]^2\right\} \\
&\quad = mn\left\{N\left[\sum i^4-4\frac{N+1}{2}\sum i^3+6\frac{(N+1)^2}{4}\sum i^2-4\frac{(N+1)^3}{8}\right.\right. \\
&\qquad \left.\left.\times\sum i+\frac{N(N+1)^4}{16}\right]-\left[\frac{N(N^2-1)}{12}\right]^2\right\}
\end{aligned}$$

Using the following relations, which can be easily proved by induction,

$$\sum_{i=1}^{N} i^3 = \left[\frac{N(N+1)}{2}\right]^2$$

$$\sum_{i=1}^{N} i^4 = \frac{N(N+1)(2N+1)(3N^2+3N-1)}{180}$$

and simplifying, the desired result is

$$\text{var}(M_N) = \frac{mn(N+1)(N^2-4)}{180} \tag{2.3}$$

The exact null probability distribution of M_N can be derived by enumeration in small samples. The labor is somewhat reduced by noting that since $a_i = a_{N-i+1}$, the properties of Theroem 7.3.7 apply. From Theroem 7.3.5 the distribution is symmetric about $N(N^2-1)/24$ when $m = n$, but the symmetry property does not hold for unequal sample sizes. Exact critical values are tabled in Laubscher, Steffens, and DeLange (1968). For larger sample sizes, the normal approximation can be used with the moments in (2.2) and (2.3). Under the assumption of normal populations differing only in variance, the asymptotic relative efficiency of the Mood test to the F test is $15/2\pi^2 = 0.76$.

9.3 THE FREUND-ANSARI-BRADLEY-DAVID-BARTON TESTS

In the Mood test of the last section, the deviation of each rank from its average rank was squared to eliminate the problem of positive and negative deviations balancing out. If the absolute values of these deviations are used instead to give equal weight to positive and negative deviations, the linear rank statistic is

$$A_N = \sum_{i=1}^{N} \left| i - \frac{N+1}{2} \right| Z_i = (N+1) \sum_{i=1}^{N} \left| \frac{i}{N+1} - \frac{1}{2} \right| Z_i \tag{3.1}$$

There are several variations of this test statistic in the literature, proposed mainly by Freund and Ansari (1957), Ansari and Bradley (1960), and David and Barton (1958). There seems to be some confusion over which test should be attributed to whom, but they are all essentially equivalent anyway.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The *Freund-Ansari-Bradley test* can be written as a linear rank statistic in the form

$$F_N = \sum_{i=1}^{N}\left(\frac{N+1}{2} - \left|i - \frac{N+1}{2}\right|\right)Z_i = \frac{m(N+1)}{2} - A_N \tag{3.2}$$

or

$$F_N = \sum_{i=1}^{[(N+1)/2]} iZ_i + \sum_{i=[(N+1)/2]+1}^{N} (N-i+1)Z_i \tag{3.3}$$

where $[x]$ denotes the largest integer not exceeding the value of x. Specifically the weights assigned then are 1 to both the smallest and largest observations in the combined sample, 2 to the next smallest and next largest, etc., $N/2$ to the two middle observations if N is even, and $(N+1)/2$ to the one middle observation if N is odd. Since the smaller weights are at the two extremes here, which is the reverse of the assignment for the Mood statistic, a small value of F_N would suggest that the X population has larger dispersion. The appropriate rejection regions for the scale-model alternative

$$H_S\colon F_{Y-M}(x) = F_{X-M}(\theta x) \qquad \text{for all } x \text{ and some } \theta > 0, \theta \neq 1$$

are then

Subclass of alternatives	*Rejection region*	*P value*
$\theta > 1$	$F_N \leqslant k_1$	$P(F_N \leqslant f \mid H_0)$
$\theta < 1$	$F_N \geqslant k_2$	$P(F_N \geqslant f \mid H_0)$
$\theta \neq 1$	$F_N \leqslant k_3$ or $F_N \geqslant k_4$	2(smaller of above)

The fact that this test is consistent for these subclasses of alternatives will be shown later in Section 9.7.

To determine the critical values for rejection, the exact null distribution of F_N could be found by enumeration. From Theorem 7.3.6, we note that the null distribution of F_N is symmetric about its mean if N is even. A recursion relation may be used to generate the null distribution systematically. For a sequence of $m+n=N$ letters occurring in a particular order, let $r_{m,n}(f)$ denote the number of distinguishable arrangements of m X and n Y letters such that the value of the F_N statistic is the number f, and let $p_{m,n}(f)$ denote the corresponding probability. A sequence of N letters is formed by adding a letter to each sequence of $N-1$ letters. If $N-1$ is even (N odd), the extra score will be $(N+1)/2$, so that f will be increased by $(N+1)/2$ if the new letter is X and be unchanged if Y. If $N-1$ is odd, the extra score will be $N/2$. Therefore we have the relations

$$N \text{ odd:} \qquad r_{m,n}(f) = r_{m-1,n}\left(f - \frac{N+1}{2}\right) + r_{m,n-1}(f)$$

$$N \text{ even:} \qquad r_{m,n}(f) = r_{m-1,n}\left(f - \frac{N}{2}\right) + r_{m,n-1}(f)$$

These can be combined in the single recurrence relation

$$r_{m,n}(f) = r_{m-1,n}(f-k) + r_{m,n-1}(f) \qquad \text{for } k = \left[\frac{N+1}{2}\right]$$

Then in terms of the probabilities, the result is

$$p_{m,n}(f) = r_{m,n}(f) \Big/ \binom{m+n}{m}$$

$$(m+n)p_{m,n}(f) = mp_{m-1,n}(f-k) + np_{m,n-1}(f)$$

which is the same form as (6.6.14) and (8.2.2) for the Mann-Whitney and Wilcoxon rank-sum tests, respectively. Tables of the null probability distributions for $N \leqslant 20$ are available in Ansari and Bradley (1960).

For larger sample sizes the normal approximation to the distribution of F_N can be used. The exact mean and variance are easily found by applying the results of Theorem 7.3.2 to F_N in the forms of (3.3) and (3.2) as follows, where $x = (N+1)/2$.

$$NE(F_n) = m\left[\sum_{i=1}^{[x]} i + \sum_{i=[x]+1}^{N} (N-i+1)\right] = m\left[\sum_{i=1}^{[x]} i + \sum_{j=1}^{N-[x]} j\right]$$

$$N \text{ even:} \qquad E(F_N) = 2m\sum_{i=1}^{N/2} \frac{i}{N} = \frac{m(N+2)}{4}$$

$$N \text{ odd:} \qquad E(F_N) = \frac{m\left[2\sum_{i=1}^{(N-1)/2} i + \frac{N+1}{2}\right]}{N} = \frac{m(N+1)^2}{4N}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$
\begin{aligned}
\operatorname{var}(F_N) &= \operatorname{var}(A_N) \\
&= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2-\left(\sum_{i=1}^{N}\left|i-\frac{N+1}{2}\right|\right)^2\right] \\
&= \frac{mn}{N^2(N-1)}\left\{\frac{N^2(N^2-1)}{12}-\left[\frac{N}{m}E(A_N)\right]^2\right\} \\
&= \frac{mn}{N^2(N-1)}\left\{\frac{N^2(N^2-1)}{12}-\left[\frac{N(N+1)}{2}-\frac{N}{m}E(F_N)\right]^2\right\}
\end{aligned}
$$

N even: $$\operatorname{var}(F_N) = \frac{mn}{N^2(N-1)}\left[\frac{N^2(N^2-1)}{12}-\left(\frac{N^2}{4}\right)^2\right] = \frac{mn(N^2-4)}{48(N-1)}$$

N odd: $$\operatorname{var}(F_N) = \frac{mn}{N^2(N-1)}\left[\frac{N^2(N^2-1)}{12}-\left(\frac{N^2-1}{4}\right)^2\right] = \frac{mn(N+1)(N^2+3)}{48N^2}$$

Collecting these results, we have

$$
\begin{array}{ll}
N \text{ even} & N \text{ odd} \\
E(F_N) = m(N+2)/4 & E(F_N) = m(N+1)^2/4N \\
\operatorname{var}(F_N) = \dfrac{mn(N^2-4)}{48(N-1)} & \operatorname{var}(F_N) = \dfrac{mn(N+1)(N^2+3)}{48N^2}
\end{array}
\tag{3.4}
$$

Another test which is almost identical is generally attributed to David and Barton (1958). This test also assigns symmetric integer weights but in the reverse order. That is, scores are given starting from the middle with 1 for N even, and 0 for N odd, and going out in both directions. The *David-Barton test* can be written as a linear rank statistic as

$$B_N = \sum_{i=1}^{[(N+1)/2]} \left(\left[\frac{N+2}{2} \right] - i \right) Z_i + \sum_{i=[(N+1)/2]+1}^{N} \left(i - \left[\frac{N+1}{2} \right] \right) Z_i \tag{3.5}$$

For N even, B_N and F_N have the exact same set of weights (but rearranged), and therefore the means and variances are equal. But for N odd this is not true because of the difference in relative assignment of the one "odd" weight, i.e., the middle observation. B_N assigns a weight of 0 to this observation, while F_N scores it as $(N+1)/2$. The following results are easily verified from Theorem 7.3.2:

N even	N odd	
$E(B_N) = m(N+2)/4$	$E(B_N) = m(N^2-1)/4N$	(3.6)
$\text{var}(B_N) = \dfrac{mn(N^2-4)}{48(N-1)}$	$\text{var}(B_N) = \dfrac{mn(N+1)(N^2+3)}{48N^2}$	

The exact relationship between B_N and F_N is

$$F_N + B_N = m[(N+2)/2] \tag{3.7}$$

Since this relation is linear, the tests are equivalent in properties. Tables of the null distribution of B_N are given in David and Barton (1958) for $m = n \leqslant 8$.

Since these three tests, F_N, B_N, and A_N, are all linearly related, they all have equivalent properties. All are consistent against the same alternatives. The asymptotic relative efficiency of each to the F test is $6/\pi^2 = 0.608$ for normal populations differing only in scale.

9.4 THE SIEGEL-TUKEY TEST

Even simpler than the use of positive integer weights symmetric about the middle would be some arrangement of the first N integers. Since these are the weights used in the Wilcoxon rank-sum test W_N for location, tables of the null probability distribution would then be

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

readily available. Siegel and Tukey (1960) proposed a rearrangement of the first N positive integers as weights which does provide a statistic sensitive to differences in scale. The rearrangement for N even is

i	1	2	3	4	5	$\cdots$	$N/2$*	$\cdots$	$N-4$	$N-3$	$N-2$	$N-1$	N
a_i	1	4	5	8	9	$\cdots$	N	$\cdots$	10	7	6	3	2

*If $N/2$ is odd, $i = (N/2) + 1$ here.

and if N is odd, the middle observation in the array is thrown out and the same weights used for the reduced N. This rearrangement achieves the desired symmetry in terms of sums of pairs of adjacent weights, although the weights themselves are not exactly symmetric. Since the weights are smaller at the extremes, we should reject the null hypothesis in favor of an alternative that the X's have the greater variability when the linear rank statistic is small.

In the symbolic form of a linear rank statistic, the *Siegel-Tukey test statistic* is

$$S_N = \sum_{i=1}^{N} a_i Z_i$$

where

$$a_i = \begin{cases} 2i & \text{for } i \text{ even}, 1 < i \leqslant N/2 \\ 2i - 1 & \text{for } i \text{ odd}, 1 \leqslant i \leqslant N/2 \\ 2(N-i) + 2 & \text{for } i \text{ even}, N/2 < i \leqslant N \\ 2(N-i) + 1 & \text{for } i \text{ odd}, N/2 < i \leqslant N \end{cases} \tag{4.1}$$

Since the probability distribution of S_N is the same as that of the Wilcoxon rank-sum statistic W_N, the moments are also the same:

$$E(S_N) = \frac{m(N+1)}{2} \qquad \text{var}(S_N) = \frac{mn(N+1)}{12} \tag{4.2}$$

To find critical values of S_N, tables of the distribution of W_N may be used, like that given in Table J of the Appendix for $m \leqslant n \leqslant 10$.

The asymptotic relative efficiency of the Siegel-Tukey test is equivalent to that of the tests F_N, B_N and A_N, because of the following relations. With N even, let S'_N be a test with weights constructed in the same manner as for S_N but starting at the right-hand end of the array, as displayed in Table 4.1 for $N/2$ even.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 4.1 Weights for Siegel-Tukey test

		i						
Test	*Weights*	*1*	*2*	*3*	*4*	*5*	$\cdots$	$N/2$
S_N	a_i	1	4	5	8	9	...	N
S'_N	a'_i	2	3	6	7	10	...	$N-1$
$S_N+S'_N$	$a_i+a'_i$	3	7	11	15	19	...	$2N-1$
S''_N	$(a_i+a'_i+1)/4$	1	2	3	4	5	...	$N/2$
		i						
Test	*Weights*	$(N/2)+1$	$\cdots$	$N-4$	$N-3$	$N-2$	$N-1$	N
S_N	a_i	$N-1$	...	10	7	6	3	2
S'_N	a'_i	N	...	9	8	5	4	1
$S_N+S'_N$	$a_i+a'_i$	$2N-1$	...	19	15	11	7	3
S''_N	$(a_i+a'_i+1)/4$	$N/2$	...	5	4	3	2	1

If $N/2$ is odd, the weights $a_{N/2}$ and $a'_{N/2}$ are interchanged, as are $a_{(N/2)+1}$ and $a'_{(N/2)+1}$. In either case, the weights $\left(a_i + a'_i + 1\right)/4$ are equal to the set of weights for F_N when N is even, and therefore the following complete cycle of relations is established for N even:

$$S''_N = F_N = m\left(\frac{N}{2}+1\right) - B_N = \frac{m(N+1)}{2} - A_N \tag{4.3}$$

9.5 THE KLOTZ NORMAL-SCORES TEST

The *Klotz (1962) normal-scores test* for scale uses the same idea as the Mood test in that it employs as weights the squares of the weights used in the inverse-normal-scores test for location [van der Waerden test of (8.3.2)]. Symbolically, the test statistic is

$$K_N = \sum_{i=1}^{N} \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2 Z_i \tag{5.1}$$

where $\Phi(x)$ is the cumulative standard normal probability distribution. Since the larger weights are at the extremes, we again reject H_0 for large K_N for the alternative that the X population has the larger

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

spread. Tables of critical values for $N \leqslant 20$ are give in Klotz (1962). The moments are

$$E(K_N) = \frac{m}{N}\sum_{i=1}^{N}\left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2$$

$$\text{var}(K_N) = \frac{mn}{N(N-1)}\sum_{i=1}^{N}\left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^4 - \frac{n}{m(N-1)}[E(K_N)]^2$$

Since this is an asymptotically optimum test against the alternative of normal distributions differing only in variance, its ARE relative to the F test equals 1 when both populations are normal.

An asymptotically equivalent test proposed by Capon (1961) uses the expected values of the square of the normal order statistics as weights or

$$\sum_{i=1}^{N}[E(\xi_{(i)}^2)]Z_i \tag{5.2}$$

where $\xi_{(i)}$ is the ith-order statistic from a standard normal distribution. This test is the scale analog of the Terry test for location in (8.3.1). The weights are tabled in Teichroew (1956), Sarhan and Greenberg (1962) for $N \leqslant 20$, and Tietjen, Kahaner, and Beckman (1977) for $N \leqslant 50$.

9.6 THE PERCENTILE MODIFIED RANK TESTS FOR SCALE

If the T_s and B_r statistics defined in (8.3.3) are added instead of subtracted, the desired symmetry of weights to detect scale differences is achieved. When N is even and $S = R = N/2$, $T + B$ is equivalent to the David-Barton type of test. The mean and variance of the statistic for N even and $S = R$ are

$$E(T_s + B_r) = \frac{mS^2}{N} \qquad \text{var}(T_s + B_r) = \frac{mnS(4NS^2 - N - 6S^3)}{6N^2(N-1)}$$

The null distribution is symmetric for $S = R$ when $m = n$. Tables for $m = n \leqslant 6$ are given in Gibbons and Gastwirth (1966), and, as for the location problem, the normal approximation to critical values may be used for $m = n \geqslant 6$.

This scale test has a higher asymptotic relative efficiency than its full-sample counterparts for all choices of $s = r < 0.50$. The maximum ARE (with respect to s) is 0.850, which occurs for normal alternatives when $s = r = 1/8$. This result is well above the ARE of 0.76 for Mood's

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

test and the 0.608 value for the tests of Sections 9.3 and 9.4. Thus asymptotically at least, in the normal case, a test based on only the 25 percent of the sample at each of the extremes is more efficient than a comparable test using the entire sample. The normal-scores tests of Section 9.5 have a higher ARE, of course, but they are more difficult to use because of the complicated sets of scores.

9.7 THE SUKHATME TEST

A number of other tests have been proposed for the scale problem. The only other one we shall discuss in detail here is the Sukhatme test statistic. Although it is less useful in applications than the others, this test has some nice theoretical properties. The test also has the advantage of being easily adapted to the construction of confidence intervals for the ratio of the unknown scale parameters.

When the X and Y populations have or can be adjusted to have equal medians, we can assume without loss of generality that this common median is zero. If the Y's have a larger spread than the X's, those X observations which are negative should be larger than most of the negative Y observations, and the positive observations should be arranged so that most of the Y's are larger than the X's. In other words, most of the negative Y's should precede negative X's, and most of the positive Y's should follow positive X's. Using the same type of indicator variables as for the Mann-Whitney statistic (6.6.2), we define

$$D_i = \begin{cases} 1 & \text{if } Y_j < X_i < 0 \text{ or } 0 < X_i < Y_j \\ 0 & \text{otherwise} \end{cases}$$

and the *Sukhatme test* statistic (Sukhatme, 1957) is

$$T = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij} \tag{7.1}$$

The parameter relevant here is

$$\begin{aligned} p &= P(Y < X < 0 \text{ or } 0 < X < Y) \\ &= \int_{-\infty}^{0} \int_{-\infty}^{x} f_Y(y) f_X(x)\, dy\, dx + \int_{0}^{\infty} \int_{x}^{\infty} f_Y(y) f_X(x)\, dy\, dx \\ &= \int_{-\infty}^{0} F_Y(x)\, dF_X(x) + \int_{0}^{\infty} [1 - F_Y(x)]\, dF_X(x) \\ &= \int_{-\infty}^{0} [F_Y(x) - F_X(x)]\, dF_X(x)) + \int_{0}^{\infty} [F_X(x) - F_Y(x)]\, dF_X(x) + 1/4 \end{aligned} \tag{7.2}$$

Then the null hypothesis of identical populations has been parameterized to H_0: $p = 1/4$, and T/mn is an unbiased estimator of p since

$$E(T) = mnp$$

By redefining the parameters p, p_1, and p_2 of the Mann-Whitney statistic as appropriate for the present indicator variables D_{ij}, the variance of T can be expressed as in (6.6.10) and (6.6.11). The probabilities relevant here are

$$\begin{aligned} p_1 &= P[(Y_j < X_i < 0 \text{ or } 0 < X_i < Y_j) \cap (Y_k < X_i < 0 \text{ or } 0 < X_i < Y_k)] \\ &= P[(Y_j < X_i < 0 \text{ or } Y_k < X_i < 0)] + P[(Y_j > X_i > 0) \cap (Y_k > X_i > 0)] \\ &= \int_{-\infty}^{0} [F_X(x)]^2 \, dF_X(x) + \int_{0}^{\infty} [1 - F_Y(x)]^2 \, dF_X(x) \end{aligned} \quad (7.3)$$

$$\begin{aligned} p_2 &= P[(Y_j < X_i < 0 \text{ or } 0 < X_i < Y_j) \cap (Y_j < X_h < 0 \text{ or } 0 < X_h < Y_k)] \\ &= P[(Y_j < X_i < 0) \cap (Y_j < X_k < 0)] + P[(Y_j > X_i > 0) \cap (Y_j > X_k > 0)] \\ &= \int_{-\infty}^{0} [1/2 - F_X(y)]^2 \, dF_Y(y) + \int_{0}^{\infty} [F_X(y) - 1/2]^2 \, dF_Y(y) \end{aligned} \quad (7.4)$$

Then from (6.6.11), the variance of T is

$$\text{var}(T) = mn[p - p^2(N-1) + (n-1)p_1 + (m-1)p_2] \quad (7.5)$$

Since $E(T/mn) = p$ and $\text{var}(T/mn) \to 0$ as m, n approach infinity, the Sukhatme statistic provides a consistent test for the following cases in terms of p in (7.2) and $\varepsilon = p - 1/4$ so that

$$\varepsilon = \int_{-\infty}^{0} [F_Y(x) - F_X(x)] \, dF_X(x) + \int_{0}^{\infty} [F_X(x) - F_Y(x)] \, dF_X(x) \quad (7.6)$$

Subclass of alternatives	*Rejection region*	*P value*	
$p < 1/4$ $(\varepsilon < 0)$ $(\theta > 1)$	$T - mn/4 \leqslant k_1$	$P(T \leqslant t \mid H_0)$	
$p > 1/4$ $(\varepsilon > 0)$ $(\theta < 1)$	$T - mn/4 \geqslant k_2$	$P(T \geqslant t \mid H_0)$	(7.7)
$p \neq 1/4$ $(\varepsilon \neq 0)$ $(\theta \neq 1)$	$\lvert T - mn/4 \rvert \geqslant k_3$	2(smaller of above)	

It would be preferable to state these subclasses of alternatives as a simple relationship between $F_Y(x)$ and $F_X(x)$ instead of this integral expression for ε. Although (7.6) defines a large subclass, we are particularly interested now in the scale alternative model where $F_Y(x) = F_X(\theta x)$. Then

1. If $\theta < 1$, $F_Y(x) > F_X(x)$ for $x < 0$ and $F_Y(x) < F_X(x)$ for $x > 0$.
2. If $\theta > 1$, $F_Y(x) < F_X(x)$ for $x < 0$ and $F_Y(x) > F_X(x)$ for $x > 0$.

In both cases, the two integrands in (7.6) have the same sign and can therefore be combined to write

$$\varepsilon = \pm \int_{-\infty}^{\infty} |F_X(\theta x) - F_X(x)|\, dF_X(x) \tag{7.8}$$

where the plus sign applies if $\theta < 1$ and the minus if $\theta > 1$. This explains the statements of subclasses in terms of θ given in (7.7).

The exact null distribution of T can be found by enumeration or a recursive method similar to that for the Mann-Whitney test. The null distribution of T is not symmetric for all m and n. The minimum value of T is zero and the maximum value is

$$M = UW + (m - U)(n - W) \tag{7.9}$$

where U and W denote the number of X and Y observations respectively which are negative. The minimum and maximum occur when the X or Y variables are all clustered. Tables of the exact distribution of T are given in Laubscher and Odeh (1976) and these should be used to find critical values for small samples.

Another test statistic which could be used for this situation is

$$T' = \sum_{i=1}^{m}\sum_{j=1}^{n} D'_{ij} = M - T \qquad \text{where } D'_{ij} = \begin{cases} 1 & \text{if } X_i < Y_j < 0 \\ & \quad \text{or } 0 < Y_j < X_i \\ 0 & \text{otherwise} \end{cases} \tag{7.10}$$

where M is defined in (7.9). Then a two-sided critical region could be written as $T \leqslant t_{\alpha/2}$ or $T' \leqslant t'_{\alpha/2}$ where $t_{\alpha/2}$ and $t'_{\alpha/2}$ have respective left-tail probabilities equal to $\alpha/2$.

For larger sample sizes, U and W converge, respectively, to $m/2$ and $n/2$ and M converges to $mn/2$ while the distribution of T approaches symmetry and the normal distribution. Laubscher and Odeh (1976) showed that this approximation is quite good for m and n

larger than ten. In the null case where $F_Y(x) = F_X(x)$ for all $x,\ p = 1/4$ and $p_1 = p_2 = 1/12$. Substituting these results in (7.5) gives the null mean and variance as

$$E(T) = mn/4 \qquad \text{and} \qquad \text{var}(T) = (mn(N+7))/48$$

For moderate m and n, the distribution of

$$\frac{4\sqrt{3}(T - mn/4)}{\sqrt{mn(N+7)}} \tag{7.11}$$

may be well approximated by the standard normal.

Ties will present a problem for the T test statistic whenever an $X_i = Y_j$, or $X_i = 0$, or $Y_j = 0$. The T statistic could be redefined in a manner similar to (6.6.16) so that a correction for ties can be incorporated into the expression for the null variance.

The Sukhatme test has a distinct disadvantage in application inasmuch as it cannot be employed without knowledge of both of the individual population medians M_X and M_Y. Even knowledge of the difference $M_Y - M_X$ is not enough to adjust the observations so that both populations have zero medians. Since the sample medians do converge to the respective population medians, the observations might be adjusted by subtracting the X and Y sample medians from each of the X and Y observations, respectively. The test statistic no longer has the same exact distribution, but for large sample sizes the error introduced by this estimating procedure should not be too large.

The Sukhatme test statistic can be written in the form of a linear rank statistic by a development similar to that used in Section 8.2 to show the relationship between the Wilcoxon and Mann-Whitney tests. Looking at (7.1) now, we know that for all values of i, $\sum_{j=1}^{n} D_{ij}$ is the sum of two quantities:

1. The number of values of j for which $Y_j < X_i < 0$, which is $r_{XY}(X_i) - U_i$
2. The number of values of j for which $Y_j > X_i > 0$, which is

   $$N - r_{XY}(X_i) + 1 - V_i$$

   where

   U_i is the number of X's less than or equal to X_i for all $X_i < 0$

   V_i is the number of X's greater than or equal to X_i for all $X_i > 0$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Then for $Z_i = 1$ if the ith variable in the combined array is an X and $Z_i = 0$ otherwise, we have

$$\begin{aligned} T &= \sum_{\substack{i=1 \\ X_i<0}}^{m} [r_{XY}(X_i) - U_i] + \sum_{\substack{i=1 \\ X_i>0}}^{m} [N - r_{XY}(X_i) + 1 - V_i] \\ &= \sum_{X<0} iZ_i + \sum_{X>0} (N - i + 1)Z_i - \sum U_i - \sum V_i \\ &= \sum_{X<0} iZ_i + \sum_{X>0} (N - i + 1)Z_i - \frac{U(U+1)}{2} - \frac{V(V+1)}{2} \end{aligned}$$

where $\sum_{X<0}$ indicates that the sum is extended over all values of i such that $X_i < 0$, U is the total number of X observations which are less than zero, and V is the number of X observations which are greater than zero. From this result, we can see that T is asymptotically equivalent to the Freund-Ansari-Bradley test, since as $N \to \infty$, the combined sample median will converge in probability to zero, the population median, and U and V will both converge to $m/2$, so that T converges to $F_N - m(m+2)/4$ with F_N defined as in (3.3). The test statistic is therefore asymptotically equivalent to all of the tests presented in Sections 9.3 and 9.4, and the large-sample properties are identical, including the ARE of $6/\pi^2$. Note that inasmuch as consistency is a large-sample property, the consistency of these other tests follows also from our analysis for T here.

9.8 CONFIDENCE-INTERVAL PROCEDURES

If the populations from which the X and Y samples are drawn are identical in every respect except scale, the nonparametric model of (1.2) with $M_X = M_Y = M$ is

$$F_{Y-M}(x) = F_{X-M}(\theta x) \qquad \text{for all } x \text{ and some } \theta > 0$$

Since θ is the relevant scale parameter, a procedure for finding a confidence-interval estimate of θ would be desirable. In the above model, we can assume without loss of generality that the common median M is zero. Then for all $\theta > 0$, the random variable $Y' = Y\theta$ has the distribution

$$P(Y' \leqslant y) = P(Y \leqslant y/\theta) = F_Y(y/\theta) = F_X(y)$$

and Y' and X have identical distributions. The confidence-interval estimate of θ with confidence coefficient $1 - \alpha$ should consist of all

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

values of θ for which the null hypothesis of identical populations will be accepted for the observations X_i and $Y_j\theta$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$. Using the Sukhatme test criterion of (7.1), here T denotes the number of pairs $(x_i, y_j\theta)$ for which either $y_j\theta < x_i < 0$ or $0 < x_i < y_j\theta$, or equivalently the number of positive pairs such that $x_i/y_j < \theta$. Suppose the rejection region for a two-sided test of size α based on the T criterion is to reject H_0 for $T \leqslant k_1$ or $T \geqslant k_2$. The appropriate confidence interval with coefficient $1 - \alpha$ is then

$$\left(\frac{x_i}{y_j}\right)_{(k)} < \theta < \left(\frac{x_i}{y_j}\right)_{(k')} \tag{8.1}$$

where $(x_i/y_j)_{(k)}$ and $(x_i/y_j)_{(k')}$ denote the kth and k'th smallest in an array made from only those ratios x/y which are positive. For small sample sizes, k and k' are found from the tables in Laubscher and Odeh (1976). If m and n are larger than ten, the number k can be found using the normal approximation given in (7.9); the result with a continuity correction of 0.5 is

$$k = \frac{mn}{4} + 0.5 - z_{\alpha/2}\sqrt{\frac{mn(N+7)}{48}} \tag{8.2}$$

which should be rounded down to the next smaller integer. Then k' is found from $k' = mn/2 - k + 1$ since the approximate gives symmetric endpoints to the confidence interval estimate.

One other approach to obtaining a confidence interval when there is no information about location is given later in Section 9.10.

9.9 OTHER TESTS FOR THE SCALE PROBLEM

All the tests for scale presented so far in this chapter are basically of the Mann-Whitney-Wilcoxon type, and except for the Mood and Klotz tests all are asymptotically equivalent. Other tests have been proposed—some are related to these while others incorporate essentially different ideas. A few will be summarized here even though they do not all fall within the category of linear rank statistics.

A test whose rationale is similar to the two-sample median test can be useful to detect scale differences. In two populations differing only in scale, the expected proportions of the two samples between two symmetric quantile points of the combined sample would not be equal. Since the total number of observations lying between the two quantiles is fixed by the order of the quantile, an appropriate test statistic could be the number of X observations lying between these two points.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

If these quantiles are the first and third quartiles and the sample sizes are large so that the sample quartiles approach the corresponding population parameters in the null case, then the statistic might be considered asymptotically a test for equal population interquartile ranges. The null distribution of the random variable U, the number of X observations within the sample interquartile range, is the hypergeometric distribution, and the appropriate rejection region for the alternative that the X's are more widely dispersed is $U \leqslant u_\alpha$. If $m + n = N$ is divisible by 4, so that no observations equal the sample quartile values, the distribution is

$$f_U(u) = \binom{m}{u}\binom{n}{N/2 - u} \Big/ \binom{N}{N/2} \tag{9.1}$$

This test is usually attributed to Westenberg (1948).

Rosenbaum suggests that the number of observations in the X sample which are either smaller than the smallest Y or larger than the largest Y is a reasonable test criterion for scale under the assumption that the population locations are the same. The null probability that exactly r X values lie outside the extreme values of the Y sample is

$$f_R(r) = n(n-1)\binom{m}{r} B(m + n - 1 - r,\ r + 2) \tag{9.2}$$

This result is easily verified by a combinational argument (Problem 9.9). Tables of critical values are given in Rosenbaum (1953).

Another criterion, suggested by Kamat, is based on the pooled sample ranks of the extreme X and Y observations. Let R_m and R_n denote the ranges of the X ranks and Y ranks, respectively, in the combined sample ordering. If the locations are the same, a test statistic is provided by

$$D_{m,n} = R_m - R_n + n \tag{9.3}$$

Tables of critical values are given in Kamat (1956). It should be noted that when the X sample observations all lie outside the extremes of the Y sample, we have $D_{m,n} = R + n$, where R is Rosenbaum's statistic. The performance of these two tests is discussed in Rosenbaum (1965).

These three tests, as well as the others presented earlier in this chapter, are reasonable approaches to detecting dispersion differences only when the X and Y populations have the same location. If the populations do not have the same location but some measure of location is known for each population, say the medians M_X and M_Y, these values can be subtracted from the respective X and Y sample values to form samples from the $X' = X - M_X$ and $Y' = Y - M_Y$ populations which do have equal medians (in fact, zero). Then any of the tests introduced earlier in this chapter can be performed on the X' and Y' variables. This is also true if the given data can be interpreted as deviations from some specified value or norm (as in Example 10.1, Section 9.10). In this case there is an alternative approach to testing the null hypothesis of equal scale. The absolute values of the deviations $X' = |X - M_X|$ and $Y' = |Y - M_Y|$ are themselves measures of spread for the respective populations. Each of the sample deviations x_i' and y_j' are estimates of the population deviation. If these sample deviations are arranged from smallest to largest in a single array, the arrangement of x' and y' is indicative of relative spread between the two populations. Thus any of the two-sample location tests from Chapter 8 can be used on these absolute values to test for relative scale differences. This procedure will be illustrated in Example 10.1 using the Wilcoxon rank-sum test introduced in Section 8.2.

If the observations are adjusted before performing a test, say by subtracting the respective sample medians, the tests are no longer exact or even distribution-free. In fact, Moses (1963) shows that no test based on the ranks of the observations will be satisfactory for the dispersion problem without some sort of strong restriction, like equal or known medians, for the two populations. There is one type of approach to testing which avoids this problem. Although strictly speaking it does not qualify as a rank test, rank scores are used. The procedure is to divide each sample into small random subsets of equal size and calculate some measure of dispersion, e.g., the variance, range, average deviation, for each subsample. The measures for both samples can be arranged in a single sequence in order of magnitude, keeping track of which of the X and Y samples produced the measure. A two-sample location test can then be performed on the result. For example, if m and n are both divisible by 2, random pairs could be formed and the Wilcoxon rank-sum test applied to the $N/2$ derived observations of ranges of the form $|x_i - x_j|, |y_i - y_j|$. The test statistic then is an estimate of a linear function of $P(|X_i - X_j| > |Y_i - Y_j|)$. In general, for any sample dispersion measures denoted by U and V when computed for the X and Y subsamples, respectively, the Wilcoxon rank-sum test statistic estimates a linear function of $P(U > V)$. Questions such as the best subsample size and the best type of measure of

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

dispersion remain to be answered generally. Tests of this kind are called *ranklike tests*. Their ARE depends on the sizes of the random subsets, and ranges from 0.304 to a limiting value of 0.955 when the distributions are normal.

9.10 APPLICATIONS

The Siegel-Tukey test for scale differences in Section 9.4 is the most frequently used procedure because it does not require a new set of tables. The table for the distribution of the Wilcoxon rank-sum test, given here as Table J of the Appendix, can be used. We note, however, the limitation of this test in that it can detect scale differences only when the locations are the same. The null hypothesis is H_0: $\theta = \sigma_X/\sigma_Y = 1$, and the test statistic is S_N, the sum of the weights assigned to the X sample in the pooled array, where the method of assignment of all weights for $m + n = N$ even is spelled out in (4.1). The appropriate rejection regions and the P values for $m \leqslant n \leqslant 10$ are as follows, where s denotes the observed value of the test statistic S_N.

Alternative	*Rejection region*	*P value*
$\theta = \sigma_X/\sigma_Y < 1$	$S_N \geqslant w_a$	$P(S_N \geqslant s \mid H_0)$
$\theta = \sigma_X/\sigma_Y > 1$	$S_N \leqslant w'_a$	$P(S_N \leqslant s \mid H_0)$
$\theta = \sigma_X/\sigma_Y \neq 1$	$S_N \geqslant w_{a/2}$ or $S_N \geqslant w'_{a/2}$	2(smaller of above)

For larger sample sizes, the appropriate rejection regions and P values based on the normal approximation with a continuity correction of 0.5 are as follows:

Alternative	*Rejection region*	*P value*
$\theta = \frac{\sigma_X}{\sigma_Y} < 1$	$S_N \geqslant \frac{m(N+1)}{2} + 0.5 + z_a\sqrt{\frac{mn(N+1)}{12}}$	$1 - \Phi\left[\frac{s - 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}\right]$
$\theta = \frac{\sigma_X}{\sigma_Y} > 1$	$S_N \leqslant \frac{m(N+1)}{2} - 0.5 - z_a\sqrt{\frac{mn(N+1)}{12}}$	$\Phi\left[\frac{s + 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}\right]$
$\theta = \frac{\sigma_X}{\sigma_Y} \neq 1$	Both above with $z_{a/2}$	2(smaller of above)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Example 10.1 An institute of microbiology is interested in purchasing microscope slides of uniform thickness and needs to choose between two different suppliers. Both have the same specifications for median thickness but they may differ in variability. The institute gauges the thickness of random samples of 10 slides from each supplier using a micrometer and reports the data shown below as the deviation from specified median thickness. Which supplier makes slides with a smaller variability in thickness?

Supplier X: 0.028, 0.029, 0.011, −0.030, 0.017, −0.012, −0.027, −0.018, 0.022, −0.023
Supplier Y: −0.002, 0.016, 0.005, −0.001, 0.000, 0.008, −0.005, −0.009, 0.001, −0.019

Solution Since the data given represent differences from specified median thickness, the assumption of equal locations is tenable as long as both suppliers are meeting specifications.

First, we use the Siegel-Tukey test. The data arranged from smallest to largest, with X underlined, and the corresponding assignment of weights are shown in Table 10.1. The sum of the X weights is $S_N = 60$, and Table J gives the left-tail probability for $m = 10$, $n = 10$ as $P = 0.000$. Since this is a left-tail probability, the appropriate conclusion is to reject H_0 in favor of the alternative H_1: $\sigma_X/\sigma_Y > 1$ or $\sigma_X > \sigma_Y$. The data indicate that supplier Y has the smaller variability in thickness.

The STATXACT and SAS outputs for Example 10.1 are shown below. The answers and the conclusions are the same as ours.

Table 10.1 Array of data and weights

Data	*Weight*	*Data*	*Weight*
−<u>0.030</u>	<u>1</u>	0.000	19
−<u>0.027</u>	<u>4</u>	0.001	18
−<u>0.023</u>	<u>5</u>	0.005	15
−0.019	8	0.008	14
−<u>0.018</u>	<u>9</u>	<u>0.011</u>	<u>11</u>
−<u>0.012</u>	<u>12</u>	0.016	10
−0.009	13	<u>0.017</u>	<u>7</u>
−0.005	16	<u>0.022</u>	<u>6</u>
−0.002	17	<u>0.028</u>	<u>3</u>
−0.001	20	<u>0.029</u>	<u>2</u>

Note that STATXACT provides both the exact and the asymptotic P values and so does SAS. The asymptotic P value using the STATXACT package is based on the value of the Z statistic without a continuity correction (-3.402) while the SAS package solution does use the continuity correction (-3.3639). It may be noted that at the time of this writing MINITAB does not provide any nonparametric test for scale.

```
********************************
STATXACT SOLUTION TO EXAMPLE 10.1
********************************

SIEGEL-TUKEY TEST

[  Sum of scores from population <       1 > ]


Min        Max        Mean       Std-dev     Observed    Standardized
55.00      155.0      105.0        13.23       60.00        -3.402

Asymptotic Inference:
   One-sided p-value: Pr { Test Statistic .LE. Observed }  =     0.0003
   Two-sided p-value: 2 * One-sided                         =     0.0007

Exact Inference:
   One-sided p-value: Pr { Test Statistic .LE. Observed }  =     0.0001
                    Pr {   Test Statistic .EQ. Observed }  =     0.0000
   Two-sided p-value: Pr { | Test Statistic - Mean |
                              .GE.  | Observed - Mean |     =     0.0002
   Two-sided p-value: 2*One-Sided                          =     0.0002


**************************************
SAS SOLUTION TO EXAMPLE 10.1
**************************************

Program:

DATA TIME;
INPUT GROUP Time @@;
  DATALINES;
1  0.028 1 0.029 1 0.011 1 -0.030 1 0.017 1 -0.012 1 -0.027 1 -0.018 1 0.022 1 -0.023
2 -0.002 2 0.016 2 0.005 2 -0.001 2 0.000 2  0.008 2 -0.005 2 -.009  2 0.001 2 -0.019
;
PROC NPAR1WAY ST DATA=TIME;
CLASS GROUP;
VAR TIME;
exact;
RUN;
```


MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

```
Output:

                    The NPAR1WAY Procedure

              Siegel-Tukey Scores for Variable Time
                Classified by Variable GROUP

                      Sum of       Expected        Std Dev          Mean
GROUP        N        Scores       Under H0        Under H0         Score
1           10          60.0          105.0       13.228757           6.0

2           10         150.0          105.0       13.228757          15.0


                  Siegel-Tukey Two-Sample Test

                  Statistic (S)                 60.0000

              Normal Approximation
              Z                               -3.3639
              One-Sided Pr <  Z                0.0004
              Two-Sided Pr > |Z|               0.0008

              Exact Test
              One-Sided Pr <=  S            1.028E-04
              Two-Sided Pr >= |S - Mean|    2.057E-04

               Z includes a continuity correction of 0.5.

              Siegel-Tukey One-Way Analysis


              Chi-Square          11.5714
              DF                        1
              Pr > Chi-Square      0.0007
```

Second, we use the Sukhatme test on these same data. The first step is to form separate arrays of the positive and negative deviations, with the X sample underlined.

Negatives: $-\underline{0.030}$, $-\underline{0.027}$, $-\underline{0.023}$, -0.019, $-\underline{0.018}$, $-\underline{0.012}$, -0.009, -0.005, -0.002, -0.001

Positives: 0.000, 0.001, 0.005, 0.008, $\underline{0.011}$, 0.016, $\underline{0.017}$, $\underline{0.022}$, $\underline{0.028}$, $\underline{0.029}$

We find $T = 2 + 1 = 3$ from (7.1) and $T' = 23 + 24 = 47$ from (7.10). The normal approximation is $z = -2.93$ without a continuity correction and a one-tailed P value of 0.0017. (The corrected value is $z = -2.87$ with P value $= 0.0021$.) The reader can verify the relation $T + T' = M$ where M is the maximum value of T. We note that the result is quite similar to the Siegel-Tukey test and the conclusion is the same.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The Sukhatme test is not available in SAS or STATXACT at the time of this writing.

Third, we give another alternative for equal scale. The data represent deviations $X - M$ and $Y - M$ for some common median M. If the X and Y populations are both symmetric about M, each of the differences is equally likely to be positive and negative. If, further, the variables have the same scale, then the absolute values of these deviations $|X - M|$ and $|Y - M|$ should have the same median value of zero. Note that these absolute values are themselves measures of variability. Thus we can use the Wilcoxon rank-sum test for location to measure the scale difference. Then the weights should be the ordinary ranks, i.e., the integers 1 to 20 in their natural order, and the pooled ordered data and corresponding ranks are shown in Table 10.2. The sum of the X ranks here is $W_N = 149$ and the corresponding exact P value from Table J is 0.000, a right-tail probability, which makes us conclude that the median variability measure for X is larger than the median variability measure for Y. This result, while not the same as that obtained with the Siegel-Tukey or Sukhatme tests, is consistent with both previous conclusions. This will generally be true. We note, however, that the Wilcoxon test for location on the absolute values is consistent against scale alternatives only when the data are given in the form $X - M$ and $Y - M$ or can be written this way because M is known.

The advantage of this alternative procedure is that it has a corresponding confidence interval procedure for estimation of the ratio $\theta = \sigma_X/\sigma_Y$ under the assumption of symmetry, the scale model relationship in (1.1), and the observations written in the form $X - M$ and $Y - M$ for equal medians or $X - M_X$ and $Y - M_Y$ in general.

Table 10.2 Array of absolute values of data and ranks

|Data|	*Rank*	*|Data|*	*Rank*
0.000	1	0.016	11
0.001	2.5	<u>0.017</u>	<u>12</u>
0.001	2.5	<u>0.018</u>	<u>13</u>
0.002	4	0.019	14
0.005	5.5	<u>0.022</u>	<u>15</u>
0.005	5.5	<u>0.023</u>	<u>16</u>
0.008	7	<u>0.027</u>	<u>17</u>
0.009	8	<u>0.028</u>	<u>18</u>
<u>0.011</u>	<u>9</u>	<u>0.029</u>	<u>19</u>
<u>0.012</u>	<u>10</u>	<u>0.030</u>	<u>20</u>

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The procedure is to form the mn ratios $|X_i - M_X|/|Y_j - M_Y|$ for all i, j, and arrange them from smallest to largest. The confidence interval end points are the uth smallest and uth largest among these ratios, where u is found in exactly the same manner as it was in Section 8.2 using Table J or the normal approximation.

Confidence Interval Estimate for Example 10.1 The $mn = 10(10) = 100$ ratios of absolute values $|Y - M|/|X - M|$ are shown in Table 10.3. Note that the ratios used are the reciprocals of the usual ratio and this will give an interval on σ_Y/σ_X; this is done in order to avoid division by zero. Note also that each set of sample data is written in increasing order of their absolute magnitudes so that the uth smallest and uth largest can be easily identified. For $m = 10$, $n = 10$ and confidence coefficient nearest 0.95 say, Table J gives $P = 0.022$ with rank 24 so that $u = 24$. The interval estimate is $0.06 \leqslant \sigma_Y/\sigma_X \leqslant 0.53$ with confidence coefficient $1 - 2(0.022) = 0.956$; taking the reciprocals we get $1.9 \leqslant \sigma_X/\sigma_Y \leqslant 16.7$.

We also use these data to illustrate the confidence interval estimate of σ_X/σ_Y based on the Sukhatme test procedure. Here we take only the positive ratios $(X - M)/(Y - M)$ shown in Table 10.4 and there are 50 of them. In order to avoid division by zero, we form the ratios $(Y - M)/(X - M)$ to find the interval on σ_Y/σ_X, and then take the reciprocal to obtain a confidence interval estimate of σ_X/σ_Y. The value of k from (8.2) for 95% confidence is 10.80 and we round down and use $k = 10$. The confidence interval is $0.0435 \leqslant \sigma_Y/\sigma_X \leqslant 0.571$,

Table 10.3 Ratios of absolute Values

	$\lvert X - M\rvert$									
$\lvert Y - M\rvert$	0.011	0.012	0.017	0.018	0.022	0.023	0.027	0.028	0.029	0.030
0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.001	0.09	0.08	0.06	0.06	0.05	0.04	0.04	0.04	0.03	0.03
0.001	0.09	0.08	0.06	0.06	0.05	0.04	0.04	0.04	0.03	0.03
0.002	0.18	0.17	0.12	0.11	0.09	0.09	0.07	0.07	0.07	0.07
0.005	0.45	0.42	0.29	0.28	0.23	0.22	0.19	0.18	0.17	0.17
0.005	0.45	0.42	0.29	0.28	0.23	0.22	0.19	0.18	0.17	0.17
0.008	0.73	0.67	0.47	0.44	0.36	0.35	0.30	0.29	0.28	0.27
0.009	0.82	0.75	0.53	0.50	0.41	0.39	0.33	0.32	0.31	0.30
0.016	1.45	1.33	0.94	0.89	0.73	0.70	0.59	0.57	0.55	0.53
0.019	1.73	1.58	1.12	1.06	0.86	0.83	0.70	0.68	0.66	0.63

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 10.4 Ratios $(Y - M)/(X - M)$

Negatives

	$Y-M$				
$X-M$	−0.019	−0.009	−0.005	−0.002	−0.001
−0.030	0.6333	0.3000	0.1667	0.0667	0.0333
−0.027	0.7037	0.3333	0.1852	0.0741	0.0370
−0.023	0.8261	0.3913	0.2174	0.0870	0.0435
−0.018	1.0556	0.5000	0.2778	0.1111	0.0556
−0.012	1.5833	0.7500	0.4167	0.1667	0.0833

Positives

	$Y-M$				
$X-M$	0.000	0.001	0.005	0.008	0.016
0.011	0	0.0909	04545	0.7273	1.4545
0.017	0	0.0588	0.2941	0.4706	0.9412
0.022	0	0.0455	0.2273	0.3636	0.7273
0.028	0	0.0357	0.1786	0.2857	0.5714
0.029	0	0.0345	0.1724	0.2759	0.5517

and taking the reciprocals yields $1.75 \leqslant \sigma_X/\sigma_Y \leqslant 23.0$. Note that this interval is wider than the one based on the Wilcoxon rank-sum test. This will frequently be the case.

The confidence interval procedure based on the Wilcoxon test for location can also be used when the data given are not variations from some central value and hence not measures of variability in themselves, but are from populations that can take on only positive values. Many variables fall into this category—for example, age, height, weight, income, GPA, test scores, survival times, relative efficiencies, and the like. For samples from such distributions, each of the mn ratios X_i/Y_j is itself a measure of the relative spread of the X and Y populations in the sense that it is an estimate of the range (measured from zero) of the X variable relative to the range of the Y variable if both are positive variables. In other words, we are looking at scale as measured by total spread, as opposed to spread based on a central value. Then the confidence interval endpoints are the uth smallest and the uth largest of the mn ratios X/Y, where u is found from Table J or from the normal approximation using (8.2.6). We call this the *method of positive variables* and illustrate it by Example 10.2.

Example 10.2 Two potential suppliers of streetlighting equipment, A and B, presented their bids to the city manager along with the following data as a random sample of life length in months.

A: 35, 66, 58, 83, 71
B: 46, 56, 60, 49

Test whether the life length of suppliers A and B have equal variability.

Solution Before we can test for scale, we must determine whether we can assume the locations can be regarded as equal. We will use the Wilcoxon rank-sum test. Since supplier B has fewer observations, we label it the X sample so that $m = 4$ and $n = 5$. The pooled sample array with X underlined is 35, $\underline{46}$, $\underline{49}$, $\underline{56}$, 58, $\underline{60}$, 66, 71, 83. The test statistic is $W_N = 15$ and the one-tailed exact P value from Table J is $P = 0.143$. Thus there is no reason not to assume that the locations are the same, and we use the Siegel-Tukey test for scale. The test statistic is $S_N = 24$ with a one-sided exact P value of $P = 0.206$ from Table J. We conclude that there is no difference in the scales of the A and B populations. Now we will find a confidence interval estimate of σ_B/σ_A using the method of positive variables with confidence coefficient near 0.95. From Table J with $m = 4$, $n = 5$, we find $u = 3$ for exact confidence level 0.936. The 20 ratios are shown in Table 10.5. The confidence interval estimate is $0.648 \leqslant \sigma_B/\sigma_A \leqslant 1.400$. Note that this interval includes the ratio one, as was implied by our hypothesis test. These analyses imply that there is no basis for any preference between suppliers A and B.

9.11 SUMMARY

In this chapter we have covered many different tests for the null hypothesis that the scale parameters are identical for the two populations, $\sigma_X/\sigma_Y = 1$. Each of these procedures (except for ranklike

Table 10.5 Ratios *B/A*

B \ A	35	58	66	71	83
46	1.314	0.793	0.697	0.648	0.554
49	1.400	0.845	0.742	0.690	0.590
56	1.600	0.966	0.848	0.789	0.675
60	1.714	1.034	0.909	0.845	0.723

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

tests) required some assumption about the location of the two distributions. If we can assume that the locations are the same, then each of the procedures in Sections 9.2 to 9.6 can be carried out even when the common value is unknown or unspecified. When the locations are not the same but their difference $M_X - M_Y$ is known, we can form $X' = X - (M_X - M_Y)$ and carry out the same tests on X' and Y because X' and Y now have the same location (in fact, equal to M_Y). When the locations are not the same but are both known as, say, M_X and M_Y, these values can be subtracted to form $X' = X - M_X$ and $Y' = Y - M_Y$, and all of the tests in this chapter can be carried out because the locations of X' and Y' are now the same (in fact, equal to zero). If the medians are unknown and unequal, we can estimate them from the sample medians or use ranklike tests, but these are only ad hoc procedures whose performance is unknown.

Recall from Chapter 1 that the confidence interval estimate of any parameter is the set of all values of the parameter which, if stated in the null hypothesis, would be accepted at the α level that corresponds to one minus the confidence level. Therefore, in order to develop a procedure for finding a confidence interval estimate for $\theta = \sigma_X/\sigma_Y$, we must be able to generalize the test for $\theta = 1$ to a test for $\theta = \theta_0 \neq 1$.

1. First, assume that $M_X = M_Y$, unspecified. The natural approach would be to form $X' = X/\theta_0$ so that $\sigma_{X'}/\sigma_Y = 1$. But then $M_{X'} = M_X/\theta_0$ must be equal to M_Y which cannot be true unless $\theta_0 = 1$, a contradiction unless $M_X = M_Y = 0$.
2. Second, assume that $M_X - M_Y$ is known but not equal to zero. The natural approach would be to form $X' = [X - (M_X - M_Y)]/\theta_0$ so that $\sigma_{X'}/\sigma_Y = 1$. But then $M_{X'} = M_Y/\theta_0$ must be equal to M_Y, which cannot be true unless $\theta_0 = 1$, a contradiction unless $M_X = M_Y = 0$.
3. Third, assume that M_X and M_Y are known. The natural approach would be to form $X' = (X - M_X)/\theta_0$ and $Y' = Y - M_Y$ so that $\sigma_{X'}/\sigma_{Y'} = 1$. This makes $M_{X'} = M_{Y'} = 0$ and hence we can have a test of $\theta = \theta_0 \neq 1$.

This argument shows that M_X and M_Y must both be known in order to test the null hypothesis where $\theta_0 \neq 1$, and that we can have a corresponding confidence interval procedure only in this case. The simplest ones to use are those based on the Wilcoxon rank-sum test of the absolute values and the Sukhatme test, since tables are available in each case. The corresponding confidence interval procedures were illustrated in Example 10.1.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

If we assume only that X/θ and Y are identically distributed, we can test the null hypothesis $\theta = \theta_0 \neq 1$ and this gives us the confidence interval based on the method of postive variables. This procedure was illustrated by Example 10.2. But notice that this makes $M_X = \theta M_Y$, and hence the estimate of relative scale is based on spread about the origin and not spread about some measure of central tendency.

The asymptotic relative efficiency of each of the Freund-Ansari-Bradley-David-Barton tests of Section 9.3 is 0.608 relative to the F test for normal populations differing only in scale, is 0.600 for the continuous uniform distribution, and is 0.94 for the double-exponential distribution. The ARE for the Mood test of Section 9.2 is 0.76 for normal distributions differing only in scale. The Klotz and Capon tests of Section 5 have an ARE of 1.00 in this case. The ARE of the percentile modified rank tests for scale against the F test for normal alternatives differing only in scale reaches its maximum of 0.850 when $s = r = 1/8$.

PROBLEMS

9.1. Develop by enumeration for $m = n = 3$ the null probability distribution of Mood's statistic M_N.

9.2. Develop by enumeration for $m = n = 3$ the null probability distribution of the Freund-Ansari-Bradley statistic of (3.3).

9.3. Verify the expression given in (2.3) for $\text{var}(M_N)$.

9.4. Apply Theorem 7.3.2 to derive the mean and variance of the statistic A_N defined in (3.1).

9.5. Apply Theorem 7.3.2 to derive the mean and variance of the statistic B_N defined in (3.5).

9.6. Verify the relationship between A_N, B_N, and F_N given in (4.3) for N even.

9.7. Use the relationship in (4.3) and the moments derived for F_N for N even in (3.4) to verify your answers to Problems 9.4 and 9.5 for N even.

9.8. Use Theorem 7.3.2 to derive the mean and variance of $T_s + B_r$ for N even, $S \neq R$, where $S + R \leqslant N$.

9.9. Verify the result given in (9.2) for the null probability distribution of Rosenbaum's R statistic.

9.10. Olejnik (1988) suggested that research studies in education and the social sciences should be concerned with differences in group variability as well as differences in group means. For example, a teacher can reduce variability in student achievement scores by focusing attention and classroom time on less able students; on the other hand, a teacher can increase variability in achievement by concentrating on the students with greatest ability and letting the less able students fall farther and farther behind. Previous research has indicated that mean student achievement for classes taught by teachers with a bachelor's degree is not different from that of classes taught by teachers with a master's degree. The present study was aimed at determining whether variability in student achievement is the same for these two teacher groups. The data below are the

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

achievement scores on an examination (10 = highest possible score) given to two classes of ten students. Class 1 was taught by a teacher with a master's degree and class 2 by a teacher with a bachelor's degree. The mean score is 5 for each class. is there a difference in variability of scores?

Class 1	*Class 2*
7	3
4	6
4	7
5	9
4	3
6	2
6	4
4	8
3	2
7	6

9.11. The psychology departments of public universities in each of two different states accepted seven and nine applicants, respectively, for graduate study next fall. Their respective scores on the Graduate Record Examination are:

University X: 1200, 1220, 1300, 1170, 1080, 1110, 1130

University Y: 1210, 1180, 1000, 1010, 980, 1400, 1430, 1390, 970

The sample median and mean scores for the two universities are close to equal, so an assumption of equal location may well be justified. Use the Siegel-Tukey test to see which university has the smaller variability in scores, if either.

9.12. In industrial production processes, each measurable characteristic of any raw material must have some specified average value, but the variability should also be small to keep the characteristics of the end product within specifications. Samples of lead ingots to be used as raw material are taken from two different distributors; each distributor has a specification of median weight equal to 16.0 kg. The data below represent actual weight in kilograms.

X: 15.7, 16.1, 15.9, 16.2, 15.9, 16.0, 15.8, 16.1, 16.3, 16.5, 15.5

Y: 15.4, 16.0, 15.6, 15.7, 16.6, 16.3 16.4, 16.8, 15.2, 16.9, 15.1

(*a*) Use the deviations from specified median weight to find two different interval estimates of σ_X/σ_Y with confidence coefficient nearest 0.95.

(*b*) Use the method of positive variables to find a confidence interval estimate of the ratio X/Y of scale measured relative to zero.

9.13. Data on weekly rate of item output from two different production lines for seven weeks are as follows:

Line I: 36, 36, 38, 40, 41, 41, 42

Line II: 29, 34, 37, 39, 40, 43, 44

We want to investigate the relative variability between the two lines.

(*a*) Find a one-tailed P value using the Siegel-Tukey test and state all assumptions needed for an exact P.

(*b*) Find the one-tailed P value using the Wilcoxon procedure assuming the population medians are $M_I = M_{II} = 40$ and state all assumptions needed for an exact P.

(*c*) In (*b*), you should have found many ties. Is there another appropriate procedure for analyzing these data, one for which the ties present no problem? Explain fully and outline the procedure.

(*d*) Find a 95% confidence interval estimate of the relative scales of line 1 relative to line 2 when spread is measured from zero.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

10
Tests of the Equality of *k* Independent Samples

10.1 INTRODUCTION

The natural extension of the two-sample problem is the k-sample problem, where observations are taken under a variety of different and independent conditions. Assume that we have k independent sets of observations, one from each of k continuous populations $F_1(x), F_2(x), \ldots, F_k(x)$ where the ith random sample is of size $n_i,\ i = 1, 2, \ldots, k$ and there are a total of $\sum_{i=1}^{k} n_i = N$ observations. Note that we are again assuming the independence extends across samples in addition to within samples. The extension of the two-sample hypothesis to the k-sample problem is that all k samples are drawn from identical populations

$$H_0\colon F_1(x) = F_2(x) = \cdots = F_k(x) \qquad \text{for all } x$$

The general alternative is simply that the populations differ in some way.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The location model for the k-sample problem is that the cdf's are $F(x-\theta_1), F(x-\theta_2), \ldots, F(x-\theta_k)$, respectively, where θ_i denotes a location parameter of the ith population, frequently interpreted as the median or the treatment effect. Then the null hypothesis can be written as

$$H_0\colon \theta_1 = \theta_2 = \cdots = \theta_k$$

and the general alternative is

$$H_1\colon \theta_i \neq \theta_j \qquad \text{for at least one } i \neq j$$

In classical statistics, the usual test for this problem is the analysis-of-variance F test for a one-way classification. The underlying assumptions for this test are that the k populations are identical in shape, in fact normal, and with the same variance and therefore may differ only in location. The test of equal means or

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k$$

is, within the context of this model, equivalent to the hypothesis above of k identical populations. Denoting the observations in the ith sample by $X_{i1}, X_{i2}, \ldots, X_{in_i}$, the ith-sample mean by $\bar{X}_i = \sum_{j=1}^{n_i}(X_{ij}/n_i)$, and the grand mean by $\bar{X} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}/N)$, the classical analysis-of-variance F test statistic may be written

$$F = \frac{\dfrac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{N-k}} = \frac{\text{mean square between samples}}{\text{mean square within samples}}$$

This test statistic follows the F distribution exactly, with $k-1$ and $N-k$ degrees of freedom, under the parametric assumptions when H_0 holds. The F test is robust for equal sample sizes, but it is known to be sensitive to the assumption of equality of variances when the sample sizes are unequal.

The nonparametric techniques which have been developed for this k-sample problem require no assumptions beyond continuous populations and therefore are applicable under any circumstances, and they involve only simple calculations. We shall cover here the extensions of the two-sample median test and the control median test, the Kruskal-Wallis analysis-of-variance test, some other extensions of rank tests from the two-sample problem, and tests against ordered alternatives including comparisons with a control or standard. Finally, the chi-square test for equality of k proportions will be discussed. This latter test is applicable only to populations where the random variables are dichotomous, often called *count data*, and therefore does not fit within the basic problem of k continuous populations as defined here. However,

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

when appropriate, it is a useful k-sample technique for the hypothesis of identical populations and therefore is included in this chapter.

10.2 EXTENSION OF THE MEDIAN TEST

Under the hypothesis of identical populations, we have a single random sample of size $\sum_{i=1}^{k} n_i = N$ from the common population. The grand median δ of the pooled samples is an estimate of the median of this common population. Therefore, an observation from any of the k samples is as likely to be above δ as below it. The set of N observations will support the null hypothesis then if, for each of the k samples, about half of the observations in that sample are less than the grand sample median. A test based on this criterion is attributed to Mood (1950, pp. 398–406) and Brown and Mood (1948, 1951).

As in the two-sample case, the grand sample median δ will be defined as the observation in the pooled ordered sample which has rank $(N+1)/2$ if N is odd and any number between the two observations with ranks $N/2$ and $(N+2)/2$ if N is even. Then, for each sample separately, the observations are dichotomized according as they are less than δ or not. Define the random variable U_i as the number of observations in sample number i which are less than δ, and let t denote the total number of observations which are less then δ. Then, by the definition of δ, we have

$$t = \sum_{i=1}^{k} u_i = \begin{cases} N/2 & \text{if } N \text{ is even} \\ (N-1)/2 & \text{if } N \text{ is odd} \end{cases}$$

Letting u_i denote the observed value of U_i, we can present the calculations in the following table.

	Sample 1	*Sample 2*	$\cdots$	*Sample k*	*Total*
$< \delta$	u_1	u_2	$\cdots$	u_k	t
$\geqslant \delta$	$n_1 - u_1$	$n_2 - u_2$	$\cdots$	$n_k - u_k$	$N - t$
Total	n_1	n_2	$\cdots$	n_k	N

Under the null hypothesis, each of the $\binom{N}{t}$ possible sets of t observations is equally likely to be in the less-than-δ category, and the number of dichotomizations with this particular sample outcome is $\prod_{i=1}^{k} \binom{n_i}{u_i}$. Therefore the null probability distribution of the random

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

variables is the multivariate extension of the hypergeometric distribution, or

$$f(u_1, u_2, \ldots, u_k|t) = \binom{n_1}{u_1}\binom{n_2}{u_2}\cdots\binom{n_k}{u_k}\bigg/\binom{N}{t} \tag{2.1}$$

If any or all of the U_i differ too much from their expected value of $n_i\theta$, where θ denotes the probability that an observation from the common population is less than δ, the null hypothesis should be rejected. Generally, it would be impractical to set up join rejection regions for the test statistics $U_1, U_2, \ldots, U_k$, because of the large variety of combinations of the sample sizes $n_1, n_2, \ldots, n_k$ and the fact that the alternative hypothesis is generally two-sided for $k > 2$, as in the case of the F test. Fortunately, we can use another test criterion which, although an approximation, is reasonably accurate even for N as small as 25 if each sample consists of at least five observations. This test statistic can be derived by appealing to the analysis of goodness-of-fit tests in Chapter 4. Each of the N elements in the pooled sample is classified according to two criteria, sample number and its magnitude relative to δ. Let these $2k$ categories be denoted by (i,j), where $i = 1, 2, \ldots, k$ according to the sample number and $j = 1$ if the observation is less than δ and $j = 2$ otherwise. Denote the observed and expected frequencies for the (i,j) category by f_{ij} and e_{ij}, respectively. Then

$$\begin{aligned} f_{i1} &= u_i \\ f_{i2} &= n_i - u_i \end{aligned} \qquad \text{for } i = 1, 2, \ldots, k$$

and the expected frequencies under H_0 are estimated from the data by

$$\begin{aligned} e_{i1} &= \frac{n_i t}{N} \\ e_{i2} &= \frac{n_i(N-t)}{N} \end{aligned} \qquad \text{for } i = 1, 2, \ldots, k$$

The goodness-of-fit test criterion for these $2k$ categories from (4.2.1), Section 4.2, is then

$$\begin{aligned} Q &= \sum_{i=1}^{k}\sum_{j=1}^{2}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \\ &= \sum_{i=1}^{k}\frac{(u_i - n_i t/N)^2}{n_i t/N} + \sum_{i=1}^{k}\frac{[n_i - u_i - n_i(N-t)/N]^2}{n_i(N-t)/N} \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$= N\sum_{i=1}^{k}\frac{(u_i - n_i t/N)^2}{n_i t} + N\sum_{i=1}^{k}\frac{(n_i t/N - u_i)^2}{n_i(N-t)}$$

$$= N\sum_{i=1}^{k}\frac{(u_i - n_i t/N)^2}{n_i}\left(\frac{1}{t} + \frac{1}{N-t}\right)$$

$$= \frac{N^2}{t(N-t)}\sum_{i=1}^{k}\frac{(u_i - n_i t/N)^2}{n_i} \tag{2.2}$$

and Q has approximately the chi-square distribution under H_0. The parameters estimated from the data are the $2k$ probabilities that an observation is less then δ for each of the k samples and that it is not less then δ. But for each sample these probabilities sum to 1, and so there are only k independent parameters estimated. The number of degrees of freedom for Q is then $2k - 1 - k$, or $k - 1$. The chi-square approximation to the distribution of Q is somewhat improved by multiplication of Q by the factor $(N-1)/N$. Then the rejection region is

$$Q \in R \qquad \text{for } \frac{(N-1)Q}{N} \geqslant \chi^2_{k-1,\alpha}$$

As with the two-sample median test, tied observations do not present a problem unless there is more than one observation equal to the median, which can occur only for N odd, or if N is even and the two middle observations are equal. The conservative approach is suggested, whereby the decision is based on that resolution of ties which leads to the smallest value of Q.

Example 2.1 A study has shown that 45 percent of normal sleepers snore occasionally while 25 percent snore almost all the time. More than 300 patents have been registered in the U.S. Patent Office for devices purported to stop snoring. Three of these devices are a squeaker sewn into the back of night clothes, a tie to secure the wrists to the sides of the bed, and a chin strap to keep the mouth shut. An experiment was conducted to determine which device is the most effective in stopping snoring or at least in reducing it. Fifteen men who are habitual snorers were divided randomly into three groups to test the devices. Each man's sleep was monitored for one night by a machine that measures amount of snoring on a 100-point scale while using a device. Analyze the results shown below to determine whether the three devices are equally effective or not.

Squeaker	*Wrist tie*	*Chin strap*
73	96	12
79	92	26
86	89	33
91	95	8
35	76	78

Solution The overall sample median is 78. Since $N = 15$ is odd, we have $t = 7$ and the data are

Group	*1*	*2*	*3*
< 78	2	1	4
$\geqslant 78$	3	4	1

We calculate $Q = 3.75$ from (2.2) and $(N-1)Q/N = 3.50$. With df = 2 we find $0.10 < P < 0.25$ from Table B of the Appendix. There is no evidence that the three medians differ.

The STATXACT solution to Example 2.1 is shown below. Note that the results do not agree with ours. This is because they define the U statistics as the number of sample observations that are less than or equal to delta, rather than the number strictly less than delta as we did. This means that they define $t = (N+1)/2$ whenever N is odd, while we define $t = (N-1)/2$. The U statistics with this definition of $\leqslant 78$ are 2, 1, and 5, for groups 1, 2, and 3, respectively. The reader can verify that these values make $Q = 6.964$ as the printout shows. The difference in the answers is surprisingly large and the conclusions are not the same. The STATXACT printout also shows an exact P-value and a point probability. We discuss these after Example 2.2.

```
            **********************************
             STATXACT SOLUTION TO EXAMPLE 2.1
            ***********************************

MEDIAN TEST

Statistics based on the observed one-way layout:
        Number of groups      = 3
        Number of observaions = 15
        The overall median    =        78.00
        Observed Statistic    =        6.964

Asymptotic p-value: (based on Chi-Square distribution with 2 df )
     Pr { CH(X) .GE.       6.964 } =      0.0307
```

```
Exact p-value and point probability :
     Pr { CH(X) .GE.        6.964 } =       0.0676
     Pr { CH(X) .EQ.        6.964 } =       0.0466
```

Example 2.2 The staff of a mental hospital is concerned with which kind of treatment is most effective for a particular type of mental disorder. A battery of tests administered to all patients delineated a group of 40 patients who were similar as regards diagnosis and also personality, intelligence, and projective and physiological factors. These people were randomly divided into four different groups of 10 each for treatment. For 6 months the respective groups received (1) electroshock, (2) psychotherapy, (3) electroshock plus psychotherapy, and (4) no type of treatment. At the end of this period the battery of tests was repeated on each patient. The only type of measurement possible for these tests is a ranking of all 40 patients on the basis of their relative degree of improvement at the end of the treatment period; rank 1 indicates the highest level of improvement, rank 2 the second highest, and so forth. On the basis of these data (see Table 2.1), does there seem to be any difference in effectiveness of the types of treatment?

Table 2.1 Ranking of patients

Groups			
1	*2*	*3*	*4*
19	14	12	38
22	21	1	39
25	2	5	40
24	6	8	30
29	10	4	31
26	16	13	32
37	17	9	33
23	11	15	36
27	18	3	34
28	7	20	35

Solution We use the median test to see whether the four groups have the same location. The overall sample median is the observation with rank 20.5 since $N = 40$, and we note that $t = 20$ and $n_i t/N = 5$. The results are

Group	*1*	*2*	*3*	*4*
< 20.5	1	9	10	0
$\geqslant 20.5$	9	1	0	10

We calculate $Q = 32.8$ from (2.2) and $(N - 1)Q/N = 31.98$. From Table B with df = 3, we find $P < 0.001$ and we reject the null hypothesis of equal medians for the four groups.

The STATXACT solution for this example is shown below. The results agree with ours and always will when N is even. As with Example 2.1, the STATXACT solution shows the calculation of an exact P value and a point probability; this is exact in the sense that it is calculated using the multivariate hypergeometric distribution given in (2.1). STATXACT also provides a Monte Carlo estimate of the P-value. The reader is referred to the STATXACT manual for more details.

```
               ***********************************
                STATXACT SOLUTION TO EXAMPLE 2.2
               ***********************************


MEDIAN TEST

Statistics based on the observed one-way layout:
        Number of groups       = 4
        Number of observations = 40
        The overall median     = 20.50
        Observed Statistic     = 32.80

Asymptotic p-value: (based on Chi-Square distribution with 3 df )
     Pr { CH(X) .GE.        32.80 } =      0.0000

Exact p-value and point probability :
     Pr { CH(X) .GE.        32.80 } =      0.0000
     Pr { CH(X) .EQ.        32.80 } =      0.0000

Monte Carlo estimate of p-value :
     Pr { CH(X) .GE.        32.80 } =      0.0000
     99.00% Confidence Interval     = (0.0000, 0.0005)
```

10.3 EXTENSION OF THE CONTROL MEDIAN TEST

The control median test was presented in Chapter 6 as an alternative to the median test in the context of the two-sample problem. We showed that the control median test is simple to use, is as efficient as the median test in large samples, and is advantageous in certain experimental situations. We now present a generalization of the control median test, due to Sen (1962), to the k-sample case,

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

where the null hypothesis is that $k(\geqslant 2)$ populations are identical against the general alternative that the populations are different in some way.

Suppose that independent random samples of size $n_1, n_2, \ldots, n_k$ are available from populations 1 through k. Without any loss of generality let sample 1 be the control sample. First, we choose $q(\geqslant 1)$ fractions $0 < p_1 < p_2 < \cdots < p_q < 1$ and find the quantiles $X_1^{(1)} < X_1^{(2)} < \cdots < X_1^{(q)}$, corresponding to the fractions, from the first sample. Thus $X_1^{(i)}$ is the r_ith-order statistic of the first sample where $r_i = [n_1 p_i] + 1$ and $[x]$ denotes the largest integer not exceeding x.

The q quantiles define $(q+1)$ nonoverlapping and contiguous cells or blocks written as

$$I_j\colon (X_1^{(j)},\ X_1^{(j+1)}) \qquad \text{for } j = 0, 1, \ldots, q$$

where $X_1^{(0)} = -\infty$ and $X_1^{(q+1)} = \infty$. For the two-sample control median test we have $k = 2,\ q = 1,\ p_1 = [n_1/2] + 1$, and the test is based on the number of observations in sample 2 that belong to I_0. For $k > 2$ samples and $q \geqslant 1$ quantiles, we count for the ith sample the number of observations V_{ij} that belong to the block $I_j,\ j = 0, 1, \ldots, q$, $i = 1, 2, \ldots, k$, so that $n_i = \sum_{j=0}^{q} V_{ij}, V_{1j} = r_{j+1}$ for $j \geqslant 1$ and $V_{10} = r_1$. The generalization of the control median test is based on these counts. It may be noted that in the terminology introduced in Chapter 2, the count V_{ij} is the frequency of the jth block for the ith sample.

The derivation of the joint distribution of the counts $V_{ij}, i = 2, 3, \ldots, k,\ j = 0, 1, \ldots, q$, provides an interesting example of computing probabilities by conditioning. To this end, observe that given (conditional on) the q quantiles from the first sample, $X_1^{(1)} < X_1^{(2)} < \cdots < X_1^{(q)}$, the joint distribution of $V_{i0}, V_{i1}, \ldots, V_{iq}$, for any $i = 2, 3, \ldots, k$, is multinomial, given by

$$\frac{n_1!}{v_{i0!} v_{i1!} \cdots v_{iq!}} [F_i(X_1^{(1)}]^{v_{i0}} [F_i(X_1^{(2)}) - F_i(X_1^{(1)})]^{v_{i1}} \cdots [F_i(X_1^{(q)}) - F_i(X_1^{(q-1)})]^{v_{iq-1}} [1 - F_i(X_1^{(q)})]^{v_{iq}} \tag{3.1}$$

where $v_{iq} = n_i - (v_{i0} + v_{i1} + \cdots + v_{iq-1})$.

The desired (unconditional) joint probability distribution of $V_{i0}, V_{i1}, \ldots, V_{iq}$ can be derived by calculating the expectation of the expression in (3.1) with respect to the chosen quantiles from the first sample. The joint distribution of the q quantiles from the first sample is

$$\frac{n_1!}{(v_{10}-1)!(v_{11}-1)!\cdots(v_{1q-1}-1)!v_{1q}!}[F_1(w_1)]^{v_{10}-1}$$
$$\times[F_1(w_2)-F_1(w_1)]^{v_{11}-1}\cdots[F_1(w_q)-F_1(w_{q-1})]^{v_{1q-1}-1}[1-F_1(w_q)]^{v_{1q}}$$

where $-\infty < w_1 < w_2 < \cdots < w_q < \infty$ and $v_{1q} = n_1 - (v_{10} + v_{11} + \cdots + v_{1q-1})$. Given $X_1^{(1)}, X_1^{(2)}, \ldots, X_1^{(q)}$, the distribution of $(V_{i0}, V_{i1}, \ldots, V_{iq})$ is independent of that of $(V_{j0}, V_{j1}, \ldots, V_{jq})$for $i \neq j$. Thus we obtain the unconditional joint distribution of the V_{ij}'s as

$$P[V_{ij}{=}v_{ij}, i{=}2,3,\ldots,k; j{=}0,1,\ldots,q]$$
$$=\frac{n_1!}{v_{1q}!\prod_{j=0}^{q-1}(v_{ij}-1)!}\int\cdots_A\int\prod_{j=0}^{q-1}[F_1(w_{j+1})-F_1(w_j)]^{v_{1j}-1}[1-F_1(w_q)]^{v_{1q}}$$
$$\times\prod_{i=2}^{k}\left[\frac{n_i!}{\prod_{j=0}^{q}v_{1q}!}\prod_{j=0}^{q}[F_i(w_{j+1}-F_i(w_j)]^{v_{ij}}\right]\prod_{j=1}^{q}dF_1(w_j) \tag{3.2}$$

where the region A is defined by $-\infty = w_0 < w_1 < \cdots < w_q < w_{q+1} = \infty$. Under H_0, the unconditional joint distribution of the V_{ij}'s reduces to

$$\frac{\prod_{i=1}^{k} n_i!}{N!}\frac{v_q!}{\prod_{i=1}^{k} v_{iq}!}\prod_{j=0}^{q-1}\left[\frac{(v_j-1)!}{(v_{1j-1})!\prod_{i=2}^{k} v_{ij}!}\right] \tag{3.3}$$

where $N = \sum_{i=1}^{k} n_i$ and $v_j = \sum_{i=1}^{k} v_{ij}$ for $j = 0, 1, \ldots, q$.

As in the case of the median test, we reject the null hypothesis if any or all of the V_{ij} are too different from their expected values under the null hypothesis. An exact P value can be calculated from (3.3) corresponding to the observed values of the counts and hence we can make a decision about rejecting H_0 for a given level of significance. In practice, however, such an implementation of the test is bound to be tedious, especially for large k, q, and/or sample sizes.

Alternatively, we can use a test criterion defined as

$$Q^* = \sum_{j=0}^{q} \pi_j^{-1} \sum_{i=1}^{k} n_i \left[\frac{v_{ij}}{n_i} - \frac{v_j}{N}\right]^2$$

where $\pi_j = v_{1j}/(n_1 + 1)$ for $j = 0, 1, \ldots, q$. Massey (1951a) has considered an extension of the median test based on a similar criterion. Under the null hypothesis, the distribution of Q^* can be approximated by a chi-square distribution with $(k-1)q$ degrees of freedom, provided that N tends to infinity with $n_i/N \to c_i; 0 < c_1, \ldots, c_k < 1$ and $\pi_0, \pi_1, \ldots, \pi_q$ are all nonzero in the limit. Thus, an approximate size α test is to reject H_0 in favor of the general alternative if

$$Q^* \geqslant \chi^2_{(k-1)q,\alpha}$$

As with the median test, ties do not present any problems here except perhaps in the choice of the quantiles from the first sample. Also, the test is consistent against the general alternative under some mild conditions on the cdf's. When the distributions belong to the location family, $F_i(x) = F(x - \theta_i)$ with $H_0\colon \theta_1 = \theta_2 = \cdots = \theta_k = 0$, the test is consistent against any deviations from H_0. However, when the distributions belong to the scale family, $F_i(x) = F(\theta_i x)$ with $H_0\colon \theta_1 = \theta_2 = \cdots = \theta_k = 1$, the test is consistent provided either $q \geqslant 2$, or $q = 1$ but $\xi_1 \neq 0$, where $F_1(\xi_1) = \pi_0$.

The asymptotic power of this test as well as efficacy expressions are derived in Sen (1962). An important result is that when $q = 1$ and $\pi_0 = \pi_1 = 1/2$, the test is as efficient as the median test (ARE is one). More generally, when the same set of quantiles (i.e., the same q and the same set of p's) is used, this test is as efficient as the generalization of the median test (based on q preselected quantiles of the pooled sample) studied by Massey (1951a). Hence, when the sample sizes are large, there is no reason, on the basis of efficiency alone, to prefer one test over the other. However, as a practical matter, finding a quantile or a set of quantiles is always easier in a single sample than in the combined samples, and thus the control median test would be preferred, especially in the absence of any knowledge about the performance of the tests when sample sizes are small. Finally, with regard to a choice of q, the number of quantiles on which the test is based, there is evidence (Sen, 1962) that even though the choice depends on the class of underlying alternative specifications, $q = 1$ or 2 is usually sufficient in practice.

10.4 THE KRUSKAL-WALLIS ONE-WAY ANOVA TEST AND MULTIPLE COMPARISONS

The median test for k samples uses information about the magnitude of each of the N observations relative to a single number which is the median of the pooled samples. Many popular nonparametric k-sample tests use more of the available information by considering the relative

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

magnitude of each observation when compared with every other observation. This comparison is effected in terms of ranks.

Since under H_0 we have essentially a single sample of size N from the common population, combine the N observations into a single ordered sequence from smallest to largest, keeping track of which observation is from which sample, and assign the ranks $1, 2, \ldots, N$ to the sequence. If adjacent ranks are well distributed among the k samples, which would be true for a random sample from a single population, the total sum of ranks, $\sum_{i=1}^{N} i = N(N+1)/2$, would be divided proportionally according to sample size among the k samples. For the ith sample which contains n_i observations, the expected sum of ranks would be

$$\frac{n_i}{N}\frac{N(N+1)}{2} = \frac{n_i(N+1)}{2}$$

Equivalently, since the expected rank for any observation is the average rank $(N+1)/2$, the expected sum of ranks for n_i observations is $n_i(N+1)/2$. Denote the actual sum of ranks assigned to the elements in the ith sample by R_i. A reasonable test statistic could be based on a function of the deviations between these observed and expected rank sums. Since deviations in either direction indicate disparity between the samples and absolute values are not particularly tractable mathematically, the sum of squares of these deviations can be employed as

$$S = \sum_{i=1}^{k} \left[R_i - \frac{n_i(N+1)}{2} \right]^2 \tag{4.1}$$

The null hypothesis is rejected for large values of S.

In order to determine the null probability distribution of S, consider the ranked sample data recorded in a table with k columns, where the entries in the ith column are the n_i ranks assigned to the elements in the ith sample. Then R_i is the ith-column sum. Under H_0, the integers $1, 2, \ldots, N$ are assigned at random to the k columns except for the restriction that there be n_i integers in column i. The total number of ways to make the assignment of ranks then is the number of partitions of N distinct elements into k ordered sets, the ith of size n_i, and this is

$$N! \Big/ \prod_{i=1}^{k} n_i!$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Each of these possibilities must be enumerated and the value of S calculated for each. If $t(s)$ denotes the number of assignments with the particular value s calculated from (4.1), then

$$f_S(s) = t(s)\prod_{i=1}^{k} n_i!/N!$$

Obviously, the calculations required are extremely tedious and therefore will not be illustrated here. Tables of exact probabilities for S are available in Rijkoort (1952) for $k = 3, 4$, and 5, but only for n_i equal and very small. Critical values for some larger equal sample sizes are also given.

A somewhat more useful test criterion is a weighted sum of squares of deviations, with the reciprocals of the respective sample sizes used as weights. This test statistic, due to Kruskal and Wallis (1952), is defined as

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left[R_i - \frac{n_i(N+1)}{2} \right]^2 \tag{4.2}$$

The consistency of H is investigated in Kruskal (1952). H and S are equivalent test criteria only for all n_i equal. Exact probabilities for H are given in Table K of the Appendix for $k = 3$, all $n_i \leqslant 5$. The tables in Iman, Quade, and Alexander (1975) also cover $k = 4$, all $n_i \leqslant 4$ and $k = 5$, all $n_i \leqslant 3$ for the upper 10% of the exact distribution.

Since there are practical limitations on the range of tables which can be constructed, some reasonable approximation to the null distribution is required if a test based on the rationale if S is to be useful in application.

Under the null hypothesis, the n_i entries in column i were randomly selected from the set $\{1, 2, \ldots, N\}$. They actually constitute a random sample of size n_i drawn without replacement from the finite population consisting of the first N integers. The mean and variance of this population are

$$\mu = \sum_{i=1}^{N} \frac{i}{N} = \frac{N+1}{2}$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{[i - (N+1)/2]^2}{N} = \frac{N^2 - 1}{12}$$

The average rank sum for the ith column, $\bar{R}_i = R_i/n_i$, is the mean of this random sample, and as for any sample mean from a finite population.

$$E(\bar{R}_i) = \mu \qquad \text{var}(\bar{R}_i) = \frac{\sigma^2(N - n_i)}{n_i(N-1)}$$

Here then we have

$$E(\bar{R}_i) = \frac{N+1}{2} \qquad \text{var}(\bar{R}_i) = \frac{(N+1)(N-n_i)}{12n_i}$$

$$\text{cov}(\bar{R}_i, \bar{R}_j) = -\frac{N+1}{12}$$

Since $\bar{R}_i$ is a sample mean, if n_i is large, the Central Limit Theorem allows us to approximate the distribution of

$$Z_i = \frac{\bar{R}_i - (N+1)/2}{\sqrt{(N+1)(N-n_i)/12n_i}} \tag{4.3}$$

by the standard normal. Consequently Z_i^2 is distributed approximately as chi square with one degree of freedom. This holds for $i = 1, 2, \ldots, k$, but the Z_i are clearly not independent random variables since $\sum_{i=1}^{k} n_i \bar{R}_i = N(N+1)/2$, a constant. Kruskal (1952) showed that under H_O, if no n_i is very small, the random variable

$$\sum_{i=1}^{k} \frac{N - n_i}{N} Z_i^2 = \sum_{i=1}^{k} \frac{12n_i[\bar{R}_i - (N+1)/2]^2}{N(N+1)} = H \tag{4.4}$$

is distributed approximately as chi square with $k-1$ degrees of freedom. The approximate size α rejection is $H \geqslant \chi^2_{\alpha,k-1}$. Some other approximations to the null distribution of H are discussed in Alexander and Quade (1968) and Iman and Davenport (1976). For a discussion about the power of this test, see for example Andrews (1954).

The assumption made initially was that the populations were continuous, and this of course was to avoid the problem of ties. When two or more observations are tied within a column, the value of H is the same regardless of the method used to resolve the ties since the rank sum is not affected. When ties occur across columns, the midrank method is generally used. Alternatively, for a conservative test the ties can be broken in the way which is least conducive to rejection of H_0.

If ties to the extent t are present and are handled by the midrank method, the variance of the finite population is

$$\sigma^2 = \frac{N^2 - 1}{12} - \frac{\sum t(t^2 - 1)}{12}$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

where the sum is over all sets of ties in the population, and this expression should be used in $\text{var}(\bar{R}_i)$ for the denominator of Z_i. In this case (4.4) becomes

$$\sum_{i=1}^{k} \frac{N-n_i}{N} \left\{ \frac{\left[\bar{R}_i - \frac{N(N+1)}{2}\right]^2}{\frac{(N+1)(N-n_i)}{12n_i} - \frac{N-n_i}{n_i(N-1)} \frac{\sum t(t^2-1)}{12}} \right\}$$

$$= \sum_{i=1}^{k} \frac{12n_i\left[\bar{R}_i - \frac{N(N+1)}{2}\right]^2}{N(N+1) - \frac{N\sum t(t^2-1)}{N-1}} = \frac{H}{1 - \frac{\sum t(t^2-1)}{N(N^2-1)}} \tag{4.5}$$

The details are left as an exercise for the reader. Hence the correction for ties is simply to divide H in (4.2) by the correction factor $1 - \sum t(t^2-1)/N(N^2-1)$ where the sum is over all sets of t tied ranks.

When the null hypothesis is rejected, as in the normal theory case, one can compare any two groups, say i and j (with $1 \leqslant i<j \leqslant k$), by a *multiple comparisons procedure*. This can be done by calculating

$$Z_{ij} = \frac{|\bar{R}_i - \bar{R}_j|}{\sqrt{[N(N+1)/12](1/n_i + 1/n_j)}} \tag{4.6}$$

and comparing it to $z^* = z_{\alpha/[k(k-1)]}$, the $[\alpha/k(k-1)]$st upper standard normal quantile. If Z_{ij} exceeds z^*, the two groups are declared to be significantly different. The quantity α is called the *experimentwise error rate* or the *overall significance level*, which is the probability of at least one erroneous rejection among the $k(k-1)/2$ pairwise comparisons. Typically, one takes $\alpha = 0.20$ or even larger because we are making such a large number of statements. We note that $1-\alpha$ is the probability that all of the statements are correct. It is not necessary to make all possible comparisons, although we usually do. For convenience, we give the z^* values to three decimal places for a total of d comparisons at $\alpha = 0.20$ as follows:

d	1	2	3	4	5	6	7	8	9	10
z^*	1.282	1.645	1.834	1.960	2.054	2.128	2.189	2.241	2.287	2.326

This multiple comparisons procedure is due to Dunn (1964).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

APPLICATIONS

The Kruskal-Wallis test is the natural extension of the Wilcoxon test for location with two independent samples to the situation of k mutually independent samples from continuous populations. The null hypothesis is that the k populations are the same, but when we assume the location model this hypothesis can be written in terms of the respective location parameters (or treatment effects) as

$$H_0\colon \theta_1 = \theta_2 = \cdots = \theta_k$$

$$H_1\colon \text{At least two } \theta\text{'s differ}$$

To perform the test, all $n_1 + n_2 + \cdots + n_k = N$ observations are pooled into a single array and ranked from 1 to N. The test statistic H is easier to calculate in the following form, which is equivalent to (4.2):

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{4.7}$$

for R_i being the sum of the ranks from the ith sample. The appropriate rejection region is large values of H. The critical values or P values are found from Table K for $k = 3$, each $n_i \leqslant 5$. This statistic is asymptotically chi-square distributed with $k - 1$ degrees of freedom; the approximation is generally satisfactory except when $k = 3$ and the sample sizes are five or less. Therefore, Table B can be used when Table K cannot. When there are ties, we divide H by the correction factor.

For multiple comparisons, using (4.6), we declare treatments i and j to be significantly different in effect if

$$|\bar{R}_i - \bar{R}_j| \geqslant z^* \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{4.8}$$

If $n_i = n_j = N/k$ for all i and j, the right-hand side of (4.6) reduces to $z^* \sqrt{k(N+1)/6}$.

Example 4.1 For the experiment described in Example 2.2, use the Kruskal-Wallis test to see if there any difference in the medians of the four groups.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Solution The data are already ranked from 1 to 40 in Table 2.1 so we need only calculate the rank sums as $R_1 = 260$, $R_2 = 122$, $R_3 = 90$, $R_4 = 348$. With $n_1 = n_2 = n_3 = n_4 = 10$, we get

$$H = \frac{12}{40(41)(10)}[260^2 + 122^2 + 90^2 + 348^2] - 3(41) = 31.89$$

with 3 degrees of freedom. The P value from Table B is $P < 0.001$, so we reject the null hypothesis that the four medians are the same and therefore do a follow-up analysis by a multiple comparisons of the medians, using $\alpha = 0.20$. We have $\bar{R}_1 = 26.0$, $\bar{R}_2 = 12.2$, $\bar{R}_3 = 9.0$ and $\bar{R}_4 = 24.8$ and the right-hand side of (4.8) is 11.125. The treatments which have significantly different medians are 1 and 2, 1 and 3, 2 and 4.

The computer solutions to Example 4.1 are shown below using the MINITAB, SAS, and STATXACT packages. All of the results for H agree exactly.

```
*************************************
      MINITAB SOLUTION TO EXAMPLE 4.1
*************************************

      Kruskal-Wallis Test: C2 versus C1

          Kruskal-Wallis Test: on C2

C1          N     Median      Ave Rank         Z
1          10     25.500          26.0      1.72
2          10     12.500          12.2     -2.59
3          10      8.500           9.0     -3.59
4          10     34.500          34.8      4.47
     Overall      40                  20.5

          H = 31.89 DF = 3 P = 0.000
```

MINITAB shows the value of the test statistics as $H = 31.89$ and the asymptotic P value of 0.000 based on the chi-square approximation with 3 degree of freedom. If there had been ties in the data, MINITAB would have shown H(adjusted), which is calculated from (4.5). MINITAB also shows the median, average rank, and Z value for each group. The Z values given are calculated from (4.3). This is the standardized value of the deviation between the mean rank $\bar{R}_i$ for the ith group and its expected value $(N + 1)/2$ under the null hypothesis. The sign of the Z statistic indicates whether the mean rank is larger or smaller than expected, and the magnitude measures the relative deviation. The largest absolute Z value is 4.47, which indicates that the mean rank for group 4, which is 34.8, differs from the average rank of 20.5 more than that of any other group. And the smallest absolute Z value, 1.72, shows that the average for group 1, 26.0, differs from the average rank less than that of any other group.

Now we show the program code and the results for SAS and STATXACT.

```
*****************************
SAS SOLUTION TO EXAMPLE 4.1
*****************************
Program:

data a;
input group N;
do i=1 to N;
input battery @@;
output;
end;
cards;
1 10
19 22 25 24 29 26 37 23 27 28
2 10
14 21 2 6 10 16 17 11 18 7
3 10
12 1 5 8 4 13 9 15 3 20
4 10
39 39 40 30 31 32 33 36 34 35
;
proc nparlway wilcoxon;
class group;
run;

Output

                  The NPAR1WAY Procedure

       Wilcoxon Scores (Rank Sums) for Variable battery
          Classified by Variable group

                    Sum of     Expected      Std Dev        Mean
group      N        Scores     Under H0      Under H0       Score
1         10         260.0        205.0     32.014119       26.00
2         10         122.0        205.0     32.014119       12.20
3         10          90.0        205.0     32.014119        9.00
4         10         348.0        205.0     32.014119       34.80

         Average scores were used for ties.

                 Kruskal-Wallis Test

          Chi-Square          31.8967
          DF                        3
          Pr > Chi-Square      <.0001
```


MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

```
**********************************
STATXACT SOLUTION TO EXAMPLE 4.1
**********************************

KRUSKAL-WALLIS TEST [That the 4 populations are identically
distributed]

Statistic based on the observed data :

     T(X) = The Observed test Statistic =        31.89

Asymptotic p-value: (based on Chi-square distribution with 3 df )
     Pr { T(X) .GE.         31.89 } =       0.0000

Monte Carlo estimate of p-value :
     Pr { Statistic .GE.         31.89 } =       0.0000
     99.00% Confidence Interval  = (       0.0000,       0.0005)
```

Example 4.2 For the experiment described in Example 2.1, use the Kruskal-Wallis test to see if there is any difference in the medians of the three groups.

Solution The first step is to rank the data from 1 to 15, as shown below, where rank 1 is given to the smallest score, which indicates the most effective result.

	Squeaker	*Wrist tie*	*Chin strap*
	6	15	2
	9	13	3
	10	11	4
	12	14	1
	5	7	8
Sum	42	60	18

We calculate $\sum R^2/n = 5688/5 = 1137.6$ and $H = 12(1137.6)/15(16) - 3(16)$. Table K for $k = 3, n_1 = n_2 = n_3 = 5$ shows that $0.001 < P$ value < 0.010, so the null hypothesis of equal treatment effects should be rejected. It appears that the chin strap is the most effective device in reducing snoring since it has the smallest sum of ranks. Since the null hypothesis was rejected, we carry out a multiple comparisons procedure at the 0.20 level. We have $z^* = 1.834$ for $d = 3$ and the right-hand side of (4.8) is 5.19. The sample mean ranks are $\bar{R}_1 = 8.4$, $\bar{R}_2 = 12$, $\bar{R}_3 = 3.6$. Our conclusion is that only groups 2 and 3 have significantly different median treatment effects at the overall 0.20 significance level. Recall that our hand calculations did not lead to a rejection of the null hypothesis by the median test in Example 2.1.

The computer solutions to Example 4.2 are shown below using the SAS, STATXACT, and MINITAB packages. The results for the value of H agree exactly. The P value using the chi-square approximation is 0.012, which agrees with the outputs. Note that both STATXACT and SAS allow the user the option of computing what they call an exact P value based on the permutation distribution of the Kruskal-Wallis statistic. This can be very useful when the sample sizes are small so that the chi-square approximation could be suspect. However, the exact computation is quite time consuming even for moderate sample sizes, such as 10 as in Example 4.1. For this example, SAS finds this exact P value to be 0.0042, which makes the results seem far more significant than those inferred from either Table K or the chi-square approximation. MINITAB does not have an option to calculate an exact P value and it does not provide the correction for ties.

```
***************************
SAS SOLUTION TO EXAMPLE 4.2
***************************

              The NPAR1WAY Procedure

        Wilcoxon Scores (Rank Sums) for Variable snore
                     Classified by Variable group

                    Sum of      Expected      Std Dev          Mean
group       N       Scores      Under H0      Under H0         Score

1           5        42.0         40.0        8.164966          8.40
2           5        60.0         40.0        8.164966         12.00
3           5        18.0         40.0        8.164966          3.60

                 Kruskal-Wallis Test

       Chi-Square                             8.8800
       DF                                          2
       Asymptotic Pr >  Chi-Square            0.0118
       Exact      Pr >= Chi-Square            0.0042
```

```
***********************************
  STATXACT SOLUTION TO EXAMPLE 4.2
***********************************

KRUSKAL-WALLIS TEST [That the 3 populations are identically
distributed]

Statistic based on the observed data :

     T(X) = The Observed test Statistic =       8.880

Asymptotic p-value: (based on Chi-square distribution with 2 df )
     Pr { T(X) .GE.        8.880 } =      0.0118

Exact p-value and point probability :
     Pr { Statistic .GE.        8.880 } =      0.0042
     Pr { Statistic .EQ.        8.880 } =      0.0003


***********************************
  MINITAB SOLUTION TO EXAMPLE 4.2
***********************************
```

Kruskal-Wallis Test: C1 versus C2

```
Kruskal-Wallis Test on C1

C2          N    Median    Ave Rank        Z
1           5     79.00         8.4     0.24
2           5     92.00        12.0     2.45
3           5     26.00         3.6    -2.69
Overall    15                   8.0

H = 8.88  DF = 2  P = 0.012
```

10.5 OTHER RANK-TEST STATISTICS

A general form for any k-sample rank-test statistic which follows the rationale of the Kruskal-Wallis statistic can be developed as follows. Denote the $\sum_{i=1}^{k} n_i = N$ items in the pooled (not necessarily ordered) sample by $X_1, X_2, \ldots, X_N$, and put a subscript on the ranks as an indication of which sample the observation is a member. Thus $r_j(X_i)$ is the rank of X_i where X_i is from the jth sample, for some $j = 1, 2, \ldots, k$. The rank sum for the jth sample, previously denoted by R_j would now be denoted by $\sum_i r_j(X_i)$. Since the $r_j(X_i)$ for fixed j are a random sample of n_j numbers, for every j the sum of any monotone increasing function g of $r_j(X_i)$ should, if the null hypothesis is true, on the average be approximately equal to the average of the function for all N observations multiplied by n_j. The weighted sum of squares of these deviations provides a test criterion. Thus a general k-sample rank statistic can be written as

$$Q = \sum_{j=1}^{k} \frac{\{\sum_i g[r_j(X_i)] - n_j(\sum_{j=1}^{k} \sum_i g[r_j(X_i)])/N\}^2}{n_j} \tag{5.1}$$

For simplicity, now let us denote the set of all N values of the function $g[r_j(x_i)]$ by $a_1, a_2, \ldots, a_N$ and their mean by

$$\bar{a} = \sum_{i=1}^{N} \frac{a_i}{N} = \sum_{j=1}^{k} \sum_{i} \frac{g[r_j(x_i)]}{N}$$

It can be shown (see Hajek and Sidak, 1967, pp. 170–172) that as minimum $(n_1, n_2, \ldots, n_k) \to \infty$, under certain regularity conditions the probability distribution of

$$\frac{(N-1)Q}{\sum_{i=1}^{N} (a_i - \bar{a})^2}$$

approaches the chi-square distribution with $k-1$ degrees of freedom.

Two obvious possibilities for our function g are suggested by the scores in the two-sample location problem for the Terry (normal scores) and the van der Waerden (inverse normal scores) test statistics. Since in both these cases the scores are symmetric about zero, $\bar{a}$ is zero and the k-sample analogs are

$$T = \frac{N-1}{\sum_{i=1}^{N} [E(\xi_{(i)})]^2} \sum_{j=1}^{k} \frac{[\sum_i E(\xi_{(i)})_j]^2}{n_j}$$

$$X = \frac{N-1}{\sum_{i=1}^{N} [\phi^{-1}(\frac{i}{N+1})]^2} \sum_{j=1}^{k} \frac{[\sum_i \phi^{-1}(\frac{i}{N+1})_j]^2}{n_j}$$

The T and X tests are asymptotically equivalent as before.

Example 5.1 For the data in Example 2.2, the normal scores test is illustrated using STATXACT to see if there are any differences in the medians of the four groups.

```
******************************************************
          STATXACT SOLUTION TO EXAMPLE 5.1
******************************************************

NORMAL SCORES TEST [That the 4 populations are identically
distributed]

Statistic based on the observed data :
     The Observed Statistic =  29.27

Asymptotic p-value: (based on Chi-square distribution with 3 df )
     Pr { Statistic .GE.   29.27 } =  0.0000

Monte Carlo estimate of p-value :
     Pr { Statistic .GE.   29.27 } =  0.0000
     99.00% Confidence Interval  = { 0.0000, 0.0005}
```

The value of the T statistic is found to be 29.27 with an approximate P value close to 0 and this leads to a rejection of the null hypothesis. Recall that for these data, both the median test and the Kruskal-Wallis test also led to a rejection of the null hypothesis.

So far we have discussed the problem of testing the hypothesis that k continuous populations are identical against the general (omnibus) alternative that they differ in some way. In practice the experimenter may expect, in advance, specific kinds of departures from the null hypothesis, say in a particular direction. For example, it might be of interest to test for an increasing (or decreasing) effect of a group of treatments on some response variable. Conceptually, some of these problems can be viewed as generalizations of one-sided alternatives to the case of more than two samples. It seems reasonable to expect that we will be able to construct tests that are more sensitive (powerful) in detecting the specific departures (from the null hypothesis) than an omnibus test, like the Kruskal-Wallis test, since the latter does not utilize the prior information in a postulated ordering (the omnibus tests are used to detect "any" deviations from homogeneity).

The problem of testing the null hypothesis of homogeneity against alternative hypotheses that are more specific or restricted in some manner than a global alternative (of nonhomogeneity) has been an area of active research. The seminal work of Barlow, Bartholomew, Bremner, and Brunk (1972) and the book by Robertson, Wright, and Dykstra (1988) are excellent references to this subject. We will discuss some of these problems in the following sections.

10.6 TESTS AGAINST ORDERED ALTERNATIVES

Suppose we are concerned with testing the null hypothesis H_O that the populations are identical against the alternative hypothesis that the location parameters are in an increasing order,

$$H_1\colon \theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_k$$

where at least one of the inequalities is strict. This alternative would be of interest if, for example, it is expected that increasing the value of a factor would result in an increase in the value of a response. If the expected direction in H_1 is not the natural ordering, we may simply relabel the treatments so that the postulated order agrees with the natural order here. This alternative is also known as the "simple order." It may be noted that under the location model when $\theta_i < \theta_j$, the population corresponding to F_j is stochastically larger than the population corresponding to F_i.

A number of distribution-free tests for the problem of simple order are available in the literature. One way to motivate some of these tests is to note that the alternative hypothesis H_1 may be written in an expanded form as

$$\begin{array}{l} \theta_1 \leqslant \theta_2, \theta_1 \leqslant \theta_3, \ldots, \theta_1 \leqslant \theta_k \\ \qquad \theta_2 \leqslant \theta_3, \theta_2, \leqslant \theta_4, \ldots, \theta_2 \leqslant \theta_k \\ \qquad\qquad \cdots\cdots\cdots\cdots\cdots\cdots \\ \qquad\qquad\qquad \theta_{k-2} \leqslant \theta_{k-1}, \theta_{k-2} \leqslant \theta_k \\ \qquad\qquad\qquad\qquad \theta_{k-1} \leqslant \theta_k \end{array} \tag{6.1}$$

where at least one of the $k(k-1)/2$ inequalities must be strict. Hence the problem of testing H_0 against H_1 may be viewed as a collection of $k(k-1)/2$ test problems, each of which is a two-sample problem. This observation allows us to contemplate several tests for the problem at hand since we have a number of tests already available for the two-sample problem.

There are two basic questions we need to address. It is clear that all of the $k(k-1)/2$ two-sample test statistics must be of the same form, for example, Mann-Whitney, median, or control median. The first question is which of the available two-sample tests should be used and why. This is a problem we first encountered in Chapter 6, and we can use either the optimal test (from the ARE point of view) or some

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

test that is more attractive from a practical point of view. Now, after a two-sample test is chosen and $k(k-1)/2$ of these tests are performed, the next question is to decide how to combine these tests into a single final test with desirable properties.

A popular test for the ordered alternatives problem is a test proposed by Terpstra (1952) and Jonckheere (1954) independently, hereafter called the JT test. The JT test uses a Mann-Whitney statistic U_{ij} for the two-sample problem comparing samples i and j, where $i, j = 1, 2, \ldots, k$ with $i < j$, and an overall test statistic is constructed simply by adding all the U_{ij}. Thus the test statistic is

$$\begin{aligned} B &= U_{12} + U_{13} + \cdots + U_{1k} + U_{23} + U_{24} + \cdots + U_{2k} + \cdots + U_{k-1,k} \\ &= \sum_{1 \leqslant i < j \leqslant k} U_{ij} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} I(X_{ir} < X_{js}) \end{aligned}$$

where X_{ir} is the rth observation in sample i and X_{js} is the sth observation in sample j, and I is the usual indicator function. The appropriate rejection region is large values of B because if the alternative H_1 is true, observations from the jth sample will tend to be larger than the observations from the ith sample. Thus the appropriate rejection region is

$$B \geqslant B(\alpha, k, n_1, n_2, \ldots, n_k)$$

where $P[B \geqslant B(\alpha, k, n_1, n_2, \ldots, n_k)] \leqslant \alpha$ is satisfied under H_0.

The JT test is distribution free if the cdf's are all continuous. Since all $N!/(\Pi_{i=1}^{k} n_i!)$ rank assignments are equally likely under H_0, the null distribution of the test statistic B can be obtained by computing the associated value of B for each possible ranking and enumerating. The required calculations are bound to be tedious, especially for large n_i, and will not be illustrated here. However, some exact critical values have been tabulated for $k = 3$, $2 \leqslant n_1 \leqslant n_2 \leqslant n_3 \leqslant 8$; $k = 4, 5, 6$, $n_1 = n_2 = \cdots = n_k = 2(1)(6)$ and are given in Table R of the Appendix for selected α.

In practice, for larger sample sizes, it is more convenient to use an approximate test. If n_i/N tends to some constant between 0 and 1, the distribution of the random vector $(U_{12}, U_{13}, \ldots, U_{k-1,k})$ under H_0 can be approximated by a $k(k-1)/2$ -dimensional normal distribution.

From the results in Section 6.6 for the Mann-Whitney test we have $E(U_{ij}) = n_i n_j/2$ under H_0 so that

$$E_0(B) = \sum_{1 \leqslant i < j \leqslant k} \sum \frac{n_i n_j}{2} = \frac{N^2 - \sum_{i=1}^{k} n_i^2}{4} \tag{6.2}$$

The derivation of the variance of B is more involved and is left as an exercise for the reader. The result under H_0 is

$$\mathrm{var}_0(B) = \frac{N^2(2N+3) - \sum_{i=1}^{k} n_i^2(2n_i+3)}{72} \tag{6.3}$$

In view of these results, an approximate level α test based on the JT statistic is to reject H_0 in favor of H_1 if

$$B \geqslant E_0(B) + z_\alpha [\mathrm{var}_0(B)]^{1/2}$$

where z_α is the $(1-\alpha)$th quantile of the standard normal probability distribution.

Because of our assumption of continuity, theoretically there can be no tied observations within or between samples. However, ties do occur in practice, and when the number of ties is large the test should be modified in a suitable manner. When observation(s) from sample i are tied with observation(s) from sample j, we replace U_{ij} by U_{ij}^* (see the discussion of the problem of ties with the Mann-Whitney statistic in Chapter 6), defined as

$$U_{ij}^* = \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} D_{rs}$$

where

$$D_{rs} = \begin{cases} 1 & \text{if } X_{ir} < X_{js} \\ 1/2 & \text{if } X_{ir} = X_{js} \\ 0 & \text{if } X_{ir} > X_{js} \end{cases}$$

This is equivalent to replacing the ranks of the tied observations by their midranks in the combined samples i and j. The JT test in the case of ties is then based on

$$B^* = \sum_{i=1}^{k-1} \sum_{j=1}^{k} U_{ij}^* \tag{6.4}$$

Under the null hypothesis, the expectation of B^* is the same as that of B, given in (6.2). Also, the variance of B^* under H_0 is

$$\begin{aligned}\mathrm{var}_0(B^*) = (72)^{-1}\Bigg[&N(N-1)(2N+5) - \sum_{i=1}^{k} n_i(n_i-1)(2n_i+5) \\ &- \sum_1 t(t-1)(2t+5)\Bigg] + [36N(N-1)(N-2)]^{-1} \\ &\times \left[\sum_{i=1}^{k} n_i(n_i-1)(n_i-2)\right]\left[\sum_1 t(t-1)(t-2)\right] \\ &+ [8N(N-1)]^{-1}\left[\sum_{i=1}^{k} n_i(n_i-1)\right]\left[\sum_1 t(t-1)\right] \qquad (6.5)\end{aligned}$$

where $\sum_1$ denotes the summation over all distinct values among the N observations, and t denotes the number of occurrences (multiplicity) of a particular distinct value. For a proof of this result see, for example, Kendall and Gibbons (1990, pp. 95–96). When there are no ties in the data, $t = 1$ for all N observations and the expression (6.5) reduces to (6.3). Although tables are apparently not yet available for the exact null distribution of B^*, an approximate test for large sample sizes can be based on the statistic $[B^* - E_0(B^*)]/[\mathrm{var}_0(B^*)]^{1/2}$ and the standard normal distribution. Thus, in the case of ties, an approximately size α JT test is to reject H_0 in favor of H_1 if

$$B^* \geqslant E_0(B^*) + z_\alpha[\mathrm{var}_0(B^*)]^{1/2} \qquad (6.6)$$

The JT test is consistent against ordered alternatives under the assumption that n_i/N tends to some constant between 0 and 1 as N tends to infinity. The asymptotic relative efficiency and some asymptotic power comparisons with competing tests are discussed in Puri (1965).

A vast body of literature is available on related and alternative procedures for this problem. Chacko (1963), Hogg (1965), Puri (1965), Conover (1967), Shorack (1967), Johnson and Mehrotra (1971), Tryon and Hettmansperger (1973), Shirahata (1980), Fairley and Flinger (1987), and Hettmansperger and Norton (1987), among others, consider various extensions. Among these, Puri (1965) studies a generalization of the JT test by introducing a class of linear rank statistics using an arbitrary score function and any two-sample Chernoff-Savage statistic. When the expected normal scores are used, Puri's procedure is highly recommended (Berenson, 1982) when the samples are from (1) normal populations; (2) light-tailed populations of the beta family, regardless of symmetry; (3) heavy-tailed and moderately skewed populations; and (4) heavy-tailed and very skewed populations. Tyron and Hettmansperger generalize Puri's class of tests by including weighting coefficients to form linear combinations of two-sample Chernoff-Savage statistics and provide some interesting results about how to determine the optimal weighting coefficients. Chakraborti and Desu (1988*a*) adopt a similar approach using the two-sample control quantile (median) test statistics and show that their optimal test has higher ARE in certain situations.

APPLICATIONS

The JT test rejects H_0 against the ordered alternative H_1 when B is significantly large. Thus, the exact P value is

$$P(B \geqslant b \mid H_0)$$

where b is the observed value of the test statistic B. When sample sizes are moderately large, the normal approximation to the P value is given by

$$1 - \Phi\left[\frac{B - E_0(B)}{\sqrt{\operatorname{var}_0(B)}}\right]$$

where $E_0(B)$ and $\operatorname{var}_0(B)$ are given in (6.2) and (6.3).

Example 6.1 Experts have long claimed that speakers who use some sort of audiovisual aids in their presentations are much more effective in communicating with their audience. A consulting agency would like to test this claim; however, it is very difficult to find many speakers who can be regarded as (virtually) identical in their speaking capabilities, and the agency is successful in locating only 15 such speakers from a nationwide search. The speakers are randomly assigned to one of three groups. The first group of speakers were not allowed to use any audiovisual aids, the second group of speakers were allowed to use a regular overhead projector and a microphone, and the third group of speakers could use a 35-mm color slide projector together with a microphone and a tape recorder (which played prerecorded audio messages). After a certain period of time, each of the speakers made a presentation in an auditorium, on a certain issue, in front of a live audience and a selected panel of judges. The contents of their presentations were virtually the same, so that any differences in effectiveness could be attributed only to the audiovisual aids used by the speakers. The judges scored each presentation on a scale of 30 to 100, depending on their own judgment and the reaction of the audience, with larger scores denoting greater effectiveness; the scores are given below. It seems reasonable to expect that the use of audiovisual aids would have some beneficial effect and hence the median score for group 1 will be the lowest, that for group 3 the highest, and the median score for group 2 somewhere in between.

Group 1	*Group* 2	*Group* 3
74, 58, 68, 60, 69	70, 72, 75, 80, 71	73, 78, 88, 85, 76

Solution The hypotheses to be tested are H_0: $\theta_1 = \theta_2 = \theta_3$, where θ_i is the median of the ith group, against H_1: $\theta_1 \leqslant \theta_2 \leqslant \theta_3$, where at least one of the inequalities is strict. Here $k = 3$ and in order to apply the JT test the three two-sample Mann-Whitney statistics U_{12}, U_{13}, and U_{23} are needed. We find $U_{12} = 22, U_{13} = 24$, and $U_{23} = 21$ and hence $B = 67$. The exact P value for the JT test from Table R of the Appendix is $P(B \geqslant 67 | H_0) < 0.0044$. Thus H_0 is rejected in favor of H_1 at any commonly used value of α and we conclude that audiovisual aids do help in making a presentation more effective, and in fact, when all other factors are equal, there is evidence that the more audiovisual aids are used, the more effective is the presentation. Also, we have $E_0(B) = 37.5$ and $\text{var}_0(B) = 89.5833$, so that using the normal approximation, $z = 3.1168$ (without a continuity correction) and the approximate P value from Table A of the Appendix is $1 - \Phi(3.12) = 0.0009$; the approximate JT test leads to the same conclusion. The SAS and STATXACT computer solutions shown below agree exactly with ours.

```
********************************
 SAS SOLUTION TO EXAMPLE 6.1
********************************
```

Program:

```
data example;
input group score @@;
datalines;
1 74 1 58 1 68 1 60 1 69 2 70 2 72 2 75 2 80 2 71
3 73 3 78 3 88 3 85 3 76
;
proc freq data=example;
tables group*score/noprint;
exact JT;
run;
```

Output:

```
                   The FREQ Procedure

        Statistics for Table of group by score

        Jonckheere-Terpstra Test

           Statistic (JT)                  67.0000
           Z                                3.1168

        Asymptotic Test
        One-sided Pr >  Z                 0.0009
        Two-sided Pr > |Z|                0.0018

        Exact Test
        One-sided Pr >=  JT            5.259E-04
        Two-sided Pr >= |JT - Mean|       0.0011

           Sample Size = 15
```

```
**********************************
 STATXACT SOLUTION TO EXAMPLE 6.1
**********************************
```

```
JONCKHEERE-TERPSTRA TEST [That the 3 populations are identically
distributed]
Statistic based on the 15 observations:
```

```
     Mean         Std-dev     Observed(JT(x))    Standardized(JT*(x))
     37.50         9.465          67.00                   3.117

Asymptotic p-value:
    One-sided: Pr { JT*(X) .GE.         3.117 }  ~        0.0009
    Two-sided: 2 * One-sided                      =        0.0018

Exact p-values:
    One-sided: Pr {  JT*(X)  .GE.       3.117 }  =        0.0005
               Pr {  JT*(X)  .EQ.       3.117 }  =        0.0002
    Two-sided: Pr { |JT*(X)| .GE.       3.117 }  =        0.0011
```

10.7 COMPARISONS WITH A CONTROL

The case of ordered alternatives is one example of a situation where the experimenter wishes to detect, a priori, not just any differences among a group of populations, but differences only in some specific directions. We now consider another example where the alternative to the null hypothesis of homogeneity is restricted, a priori, in a specific direction.

Suppose we want to test only a partial ordering of the $(k-1)$ distributions with respect to a common distribution. This will be the situation where only a comparison between each of the distributions and the common distribution is of interest and what happens among the $(k-1)$ distributions is somewhat irrelevant. For example, in a drug screening study, it is often of interest to compare a group of treatments under development to what is currently in use (which may be nothing or a placebo), called the control, and then subject those treatments that are "better" than the control to more elaborate studies. Also, in a business environment, for example, people might consider changing their current investment policy to one of a number of newly available comparable policies provided the payoff is higher.

The alternative hypothesis of interest here is another generalization of the one-sided alternatives problem to the case of several samples and constitutes a subset of the ordered alternatives problem discussed earlier. One expects, at least intuitively, to be able to utilize the available pertinent information to construct a test which is more powerful than the tests for either the general or the ordered alternatives problem.

Without any loss of generality let F_1 be the cdf of the control population and let F_i be the cdf of the ith treatment population,

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$i = 2, 3, \ldots, k$, where $F_i = F(x - \theta_i)$ with $F_i(0) = p$, so that θ_i is the pth $(0 < p < 1)$ quantile of F_i. Our problem is to test, for a specified p, the null hypothesis that the pth treatment quantiles are equal and equal to the pth control quantile,

$$H_0\colon \theta_2 = \theta_3 = \cdots = \theta_k = \theta_1$$

against the one-sided alternative hypothesis,

$$H_1\colon \theta_2 \geqslant \theta_1,\ \theta_3 \geqslant \theta_1, \ldots, \theta_k \geqslant \theta_1$$

where at least one of the inequalities is strict. As pointed out earlier, when $\theta_i > \theta_1$, F_i is stochastically larger than F_1. In the literature on hypothesis testing with restricted alternatives, this alternative is known as the simple-tree alternative. Miller (1981), among others, calls this a “many-one” problem.

In some cases it might be of interest to test the alternative in the opposite direction, $\theta_2 \leqslant \theta_1,\ \theta_3 \leqslant \theta_1, \ldots, \theta_k \leqslant \theta_1$, where at least one of the inequalities is strict. The tests we discuss can easily be adapted for this case.

Our approach in the many-one problem is basically similar to that in the ordered alternatives problem in that we view the testing problem as a collection of $(k-1)$ subtesting problems $H_{0i}\colon \theta_i = \theta_1$ against $H_{1i}\colon \theta_i \geqslant \theta_1$ for $i = 2, 3, \ldots, k$. Thus, a distribution-free test for testing H_0 against H_1 is obtained in two steps. First an appropriate one-sample test for the ith subtesting problem is selected and then $(k-1)$ of these tests are combined in a suitable manner to produce an overall test.

Before considering specific tests we would like to make a distinction between the cases (i) where sufficient prior knowledge about the control is at hand so that the control population may be assumed to be known (in this case the control is often called a standard), at least to the extent of the parameter(s) of interest, and (ii) where no concrete knowledge about the control group is available. These two cases will be treated separately and the test procedures for these cases will be somewhat different.

CASE (I): θ_1 KNOWN

First consider the case of testing H_0 where θ_1 is either known or specified in advance, against the alternative H_1. In this case, the subtesting problem, for every $i = 2, 3, \ldots, k$, is a one-sample problem and therefore one of several available distribution-free tests can be used in step one. For example, if the only reasonable assumption about the parent distributions is continuity, we would use the sign test. On the other hand, if it can be assumed that the underlying distributions are symmetric about their respective preselected quantiles, we may want to use the Wilcoxon signed rank test. For simplicity and ease of presentation we detail only the tests based on the sign test, although one can proceed in a similar manner with some other one-sample distribution-free tests. Part of this discussion is from the papers by Chakraborti and Gibbons (1991, 1992).

The usual sign test statistic for testing H_{0i} against H_{1i} is based on the total number of negative differences $X_{ij} - \theta_1$ in the ith sample,

$$V_i = \sum_{j=1}^{n_i} I(X_{ij} - \theta_1 < 0)$$

$i = 2, 3, \ldots, k$, and we reject H_{0i} in favor of H_{1i} if V_i is small. With this motivation, a simple overall test of H_0 against H_1 can be based on V, the sum of the V_i's, $i = 2,\ 3, \ldots, k$, and the rejection region consists of small values of V. One practical advantage of using V is that under H_0, V has a binomial distribution with parameters $N = \sum_{i=2}^{k} n_i$ and p. Accordingly, P values or exact critical values can be found using binomial tables for small to moderate sample sizes. For larger sample sizes, the normal approximation to the binomial distribution can be used to construct tests with significance levels approximately equal to the nominal value or to find the approximate P value.

A simple modification of this sum test when the sample sizes are quite different is to use $V^* = \sum_{i=2}^{k}(V_i/n_i)$ as the test statistic since the V_i's, and hence V, may be quite sensitive to unequal sample sizes. The exact and/or approximate test can be implemented as before, although for the exact test we no longer have the convenience of using tables of the binomial distribution. The details are left as an exercise for the reader (see, for example, Chakraborti and Gibbons, 1992).

An alternative test for this problem may be obtained by applying the union-intersection principle (Roy, 1953). Here the null hypothesis H_0 is the intersection of the $(k-1)$ subalternative hypotheses H_{1i}: $\theta_i \geqslant \theta_1$. Thus, H_0 should be rejected if and only if at least one of the subnull hypotheses H_{0i} is rejected and the latter event takes place if the smallest of the statistics $V_2, V_3, \ldots, V_k$ is too small. In other words, an overall test may be based on

$$V^+ = \min\left(\frac{V_2}{n_2}, \frac{V_3}{n_3}, \ldots, \frac{V_k}{n_k}\right) \tag{7.1}$$

and one should reject H_0 in favor of H_1 if V^+ is small. The test based on V^+ is expected to be more sensitive than the sum test since a rejection of any of the $(k-1)$ subnull hypotheses here would lead to a rejection of the overall null hypothesis.

The exact distribution of V^+ can be obtained using the fact that $V_2, V_3, \ldots, V_k$ are mutually independent and under H_0 each V_i has a binomial distribution with parameters n_i and p. Thus

$$P(V^+ \leqslant v|H_0) = 1 - \prod_{i=2}^{k}\left[1 - \sum_{j=0}^{[n_i v]}\binom{n_i}{j}p^j(1-p)^{n_i-j}\right] \tag{7.2}$$

where v is a fraction between 0 and 1 and $[x]$ denotes the greatest integer not exceeding x. The exact null distribution can be used to calculate exact P values for small sample sizes. When sample sizes are large, it is more convenient to use the normal approximation to the binomial distribution to calculate approximate P values.

CASE (II): θ_1 UNKNOWN

When θ_1 is unknown, we use the same general idea of first choosing a suitable test for the ith subtesting problem (which in the present case is a two-sample problem), $i = 2, 3, \ldots, k$, and then combining the $(k-1)$ test statistics to construct an overall test statistic. However, as might be expected, the details are more involved, because the statistics to be combined are now dependent.

To study our tests in this case consider, for the ith subtesting problem, the following "i to 1" statistics

$$W_i = \sum_{j=1}^{n_i} I(X_{ij} < T)$$

where T is a suitable estimate of θ_1. By analogy with the one-sample case the quantity W_i can be called a two-sample sign statistic, which allows us to consider direct extensions of our earlier procedures to the present case. It may be noted that if in fact T is a sample order statistic (for example, when θ_1 is the median of F_1, T should be the median of the sample from F_1), W_i is simply the placement of T among the observations from the ith sample. We have seen that the distribution of W_i does not depend on F_i under H_0, and therefore any test based on the W's is a distribution-free test.

Now consider some tests based on a combination of the W's. Again, as in case (i), we can use the sum of the W's for a simple overall

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

test and reject H_0 in favor of H_1 if W is small. This test has been proposed and studied by Chakraborti and Desu (1988b) and will be referred to as the CD test. The exact (unconditional) distribution of the sum statistic W is obtained by noting that the exact distribution of W is simply the expectation of the joint distribution of the W_i's, $i = 2, 3, \ldots, k$, with respect to T and that conditional on T, the W_i's are independent binomial random variables with parameter n_i and $F_i(T)$ This yields, for $w = 0, 1, \ldots, (N - n_1)$

$$P[W = w] = \sum \int_{-\infty}^{\infty} \prod_{i=2}^{k} \binom{n_i}{a_i} [F_i(t)]^{a_i} [1 - F_i(t)]^{n_i - a_i} \, dF_T(t) \qquad (7.3)$$

where the sum is over all $a_i = 0, 1, \ldots, n_i, i = 2, 3, \ldots, k$, such that $a_2 + a_3 + \cdots + a_k = w$.

Under the null hypothesis the integral in (7.3) reduces to a complete beta integral and the exact null distribution of W can be enumerated. However, a more convenient closed-form expression for the null distribution of W may be obtained directly by arguing as follows. The statistic W is the total number of observations in treatment groups 2 through k that precede T. Hence the null distribution of W is the same as that of the two-sample precedence statistic with sample sizes n_1 and $N - n_1$ and this can be obtained directly from the results in Problems 2.28c and 6.10a. Thus we have, when T is the ith order statistic in the control sample,

$$P[W = w | H_0] = \left[\binom{N - i - w}{N - n_1 - w} \binom{i + w - 1}{w} \right] \bigg/ \binom{N}{N - n_1}$$

$$w = 0, 1, \ldots, N; \; i = 1, 2, \ldots, n_1$$

or equivalently,

$$P[W = w | H_0] = \frac{n_1}{N} \binom{N - n_1}{w} \binom{n_1 - 1}{i - 1} \bigg/ \binom{N - 1}{w + i - 1} \qquad (7.4)$$

Also, using the result in Problem 2.28d we have

$$E_0(W) = (N - n_1) \left(\frac{i}{n + 1} \right) \qquad (7.5)$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

and

$$\text{var}_0(W) = \frac{i(n_1 - i + 1)(N + 1)(N - n_1)}{(n_1 + 1)^2(n_1 + 2)} \tag{7.6}$$

The null distribution can be used to determine the exact critical value for a given level of significance α or to find the P value for an observed value of W.

For some practical applications and further generalizations, it is useful to derive the large sample distribution of W. We first find the large sample distribution of the $(k-1)$ dimensional random vector $(W_2, W_3, \ldots, W_k)$. It can be shown (Chakraborti and Desu, 1988*a*; Gastwrith and Wang, 1988) that the large sample distribution of this random vector can be approximated by a $(k-1)$-variate normal distribution. The result is stated below as a theorem.

Theorem 7.1 *Let W_N be the $(k-1)$-dimensional vector whose ith element is $N^{1/2}[W_{i+1}/n_{i+1} - F_{i+1}(\theta_1)], i = 1, 2, \ldots, k-1$. Suppose that, as $\min(n_1, n_2, \ldots, n_k) \to \infty$, we have $n_i/N \to \lambda_i, 0 < \lambda_i < 1$, $i = 1, 2, \ldots, k$. Also let $F_i'(\theta_1) = f_i(\theta_1)$ exist and be positive for $i = 1, 2, \ldots, k$. The random vector W_N converges in distribution to a $(k-1)$-dimensional normal distribution with mean vector* 0 *and covariance matrix Σ whose i,jth element is*

$$\sigma_{ij} = \frac{Q_i Q_j p(1-p)}{\lambda_1} + \frac{\delta_{ij} F_{i+1}(\theta_1)[1 - F_{i+1}(\theta_1)]}{\lambda_{i+1}}$$

where $Q_i = f_{i+1}(\theta_1)/f_1(\theta_1), i,j = 1, 2, \ldots, k-1$, and δ_{ij} is equal to 1 *if $i = j$ and is equal to* 0 *otherwise.*

From Theorem 7.1, the null distribution of W can be approximated by a normal distribution with mean $(N - n_1)p$ and variance $N(N - n_1)p(1-p)/n_1$. Thus, for case (ii), an approximately size α test based on the sum of the W's is to reject H_0 in favor of H_1 if

$$W \leqslant (N - n_1)p - z_\alpha \left[\frac{N(N - n_1)p(1-p)}{n_1}\right]^{1/2} \tag{7.7}$$

As we noted earlier, alternative nonparametric tests are available for our consideration in step 1, where the basic set of test statistics is chosen. For example, when θ_1 is unknown so that we choose a

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

two-sample test for the ith subtesting problem, we can employ the Mann-Whitney U statistic (or more generally any linear rank statistic) between the ith and the first sample, say U_{1i}, and combine these U's in some suitable manner for an overall test statistic. The resulting tests are distribution-free since the distribution of U_{1i} does not depend on either F_i or F_1 under H_0.

As in the case of the W's, the sum of the U's, say $W^* = \sum_{i=2}^{k} U_{1i}$, can be used for a simple overall test and such a test has been proposed and studied by Fligner and Wolfe (1982), hereafter referred to as FW test. The null distribution of W^* is the same as that of a two-sample Mann-Whitney statistic (see Section 6.6) with $m = n_1$ and $n = N - n_1$. It can be seen that the FW test rejects H_0 in favor of the simple-tree alternative if W^* is large so that the P value is in the upper tail. A choice between the tests based on the sum of U's and the sum of W's may be made on the basis of the ARE. Interestingly, when $p = 0.5$ (that is, when θ_i is the median of F_i), the ARE between these two tests can be shown to be the same as the ARE between the sign test and the signed-rank test, regardless of the underlying distribution. For example, when the underlying distribution F is normal, the ARE is 0.67, whereas when F is double exponential, the ARE is 1.33.

APPLICATIONS

The CD test rejects H_O against the simple-tree alternative H_1 when W is significantly small. Thus the P value is $P(W \leqslant w|H_0)$, where w is the observed value of the test statistic W. The exact P value can be calculated using the null distribution of W given in (7.4) by adding the individual probabilities under H_0. However, for moderately large sample sizes, the normal approximation to the P value is adequate. This can be calculated from $\Phi\left[\frac{W-E_0(W)}{\sqrt{\text{var}_0(W)}}\right]$, where $E_0(W)$ and $\text{var}_0(W)$ are given in (7.5) and (7.6), respectively.

Example 7.1 Consider again the problem in Example 6.1, where a group of speakers are compared with respect to their ability to communicate. Since the first group of speakers use no audiovisual aids, we could regard this as the control group. It is reasonable to expect that the use of audiovisual aids would have some beneficial effects in communication, and hence the scores for at least one of the groups 2

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and 3 should be greater than those for group 1. Thus we are interested in testing H_0: $\theta_3 = \theta_2 = \theta_1$ against H_1: $\theta_2 \geqslant \theta_1, \theta_3 \geqslant \theta_1$, where at least one of the inequalities is strict. In this example, θ_1 is unknown, and therefore we will use the CD test with $k = 3, n_1 = n_2 = n_3 = 5$. From the data we find $T = 68$, $W_2 = 0$, and $W_3 = 0$. Hence $W = 0$ and the exact P value from (7.4) is

$$\frac{5}{15}\left[\binom{15-5}{0}\binom{5-1}{3-1}\bigg/\binom{15-1}{0+3-1}\right] = 0.022$$

Therefore, at say $\alpha = 0.05$, there is evidence that the median scores of at least one of groups 2 and 3 is greater than that of group 1. In order to use the normal approximation we find $E_0(W) = 5$ and $\text{var}_0(W) = 5.7143$, from (7.5) and (7.6), and the approximate P value from Table A of the Appendix is $\Phi(-2.09) = 0.0183$. This is reasonably close to the exact P value even though the sample sizes are small.

For the Fligner-Wolfe test, we find $U_{12} = 22$ and $U_{13} = 24$ so that the P value is $P(W^* \geqslant 46 \mid H_0)$. To calculate this probability note that W^* is in fact the value of the Mann-Whitney statistic calculated between sample 1 (as the first sample) and samples 2 and 3 combined (as the second sample). Now using the relationship between U statistics and rank-sum statistics, it can be seen that the required probability is in fact equal to the probability that the rank-sum statistic between two samples of size $m = 5$ and $n = 10$ is at most 19 under H_0. This is found from Table J of the Appendix as 0.004. Thus, the evidence against the null hypothesis in favor of the simple tree alternative is stronger, on the basis of the FW test.

10.8 THE CHI-SQUARE TEST FOR *k* PROPORTIONS

There is one further k-sample test which should be mentioned, although it is applicable in a situation different from that of the other tests in this chapter. If the k populations are all Bernoulli distributions, the samples consist of count data. The k populations are all equivalent if the Bernoulli parameters $\theta_1, \theta_2, \ldots, \theta_k$ are all equal. Let $X_1, X_2, \ldots, X_k$ denote the numbers of successes in each of the k samples, respectively. For $i = 1, 2, \ldots, k$, X_i has the binomial distribution with parameter θ_i and number of trials n_i. A test statistic for the null hypothesis

$$H_0\colon \theta_1 = \theta_2 = \cdots = \theta_k \text{ unspecified}$$

against the alternative that the θ's are not all equal can be derived from the goodness-of-fit test exactly as in the case of the k-sample median test in Section 10.2, except that our classification criterion besides sample number is now simply success or failure for each sample instead of less than δ or not. Using the same notation as before, with θ_0 denoting the estimated common parameter under H_0 (which replaces t/N), we have

$$f_{i1} = x_i$$
$$f_{i2} = n_i - x_i$$
$$e_{i1} = n_i\theta_0$$
$$e_{i2} = n_i(1-\theta_0)$$

and the test criterion from (2.2) is

$$Q = \sum_{i=1}^{k} \frac{(x_i - n_i\theta_0)^2}{n_i\theta_0(1-\theta_0)}$$

where $\theta_0 = \sum_{i=1}^{k}(x_i/N)$, which has approximately the chi-square distribution with $k-1$ degrees of freedom. Hence the approximately size α test is to reject H_0 in favor of the alternative if $Q \geqslant \chi^2_{k-1,\alpha}$.

A simple form of the test statistic Q, useful for calculations, is

$$Q = \frac{1}{\theta_0(1-\theta_0)} \sum_{i=1}^{k} \frac{x_i^2}{n_i} - \frac{N\theta_0}{1-\theta_0} \tag{8.1}$$

Example 8.1 In a double-blind study of drugs to treat duodenal peptic ulcers, a large number of patients were divided into groups to compare three different treatments, antacid, antipepsin, and anticholinergic. Antacid has long been considered the major medication for ulcers; the latter two drugs act on different digestive juices. The number of patients in the groups and the percent who benefited from that treatment are shown below. Does there appear to be a difference in beneficial effects of these three treatments?

Treatment	*Number*	*Percent benefited*
Antacid	40	55
Anticholinergic	60	70
Antipepsin	75	84

Solution We must first calculate the number of patients benefited by each drug as $x_1 = 40(0.55) = 22, x_2 = 60(0.70) = 42, x_3 = 75(0.84) = 63$;

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

then the estimate of the common probability of benefit is $\theta_0 = (22 + 42 + 63)/175 = 0.726$ and

$$Q = \frac{1}{0.726(0.274)}\left(\frac{22^2}{40} + \frac{42^2}{60} + \frac{63^2}{75}\right) - \frac{175(0.726)}{0.274} = 10.97$$

with 2 degrees of freedom. Table B of the Appendix shows that $P < 0.001$, so we conclude that the probabilities of benefit from each drug are not equal.

10.9 SUMMARY

In this chapter we have been concerned with data consisting of k mutually independent random samples from k populations where the null hypothesis of interest is that the k populations are identical. Actual measurements are not required to carry out any of the tests.

When the location model is appropriate and the alternative is that the locations are not all the same, the median test extension, the Kruskal-Wallis, Terry and van der Waerden tests are all appropriate. The median test uses less information than the others and therefore may be less powerful. Further, exact P values require calculations based on the multivariate extension of the hypergeometric distribution and this is quite tedious. As in the two-sample case, the median test is primarily of theoretical interest. On the other hand, the Kruskal-Wallis test is simple to use and quite powerful. Tables of the exact distribution are available and the chi-square approximation is reasonably accurate for moderate sample sizes.

All of the tests are quicker and easier to apply than the F test and may perform better if the F test assumptions are not satisfied. Further, as in the parametric setting, nonparametric methods of multiple comparisons can be used in many cases to determine which pairs of population medians differ significantly; see, e.g., Miller (1966, 1981). The advantage of a multiple comparisons procedure over separate pairwise comparisons is that the significance level is the overall level, the probability of a Type I error in all of the conclusions reached.

If the alternative states a distinct complete ordering of the medians, the Jonckheere-Terpstra test is appropriate and exact P values can be obtained. We also discuss tests where the alternative states an ordering of medians with respect to a control group only, where the control median may be either known or unknown.

The asymptotic relative efficiency of the Kruskal-Wallis test compared to the normal theory test is at least 0.864 for any continuous

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

distribution and is 0.955 for the normal distribution. The Terry and van der Waerden tests should have an ARE of 1 under these circumstances, since they are asymptotically optimum for normal distributions. The ARE of the median test is only $2/\pi = 0.637$ relative to the F test for normal populations with equal variances, and 2/3 relative to the Kruskal-Wallis test, in each case for normal distributions. For further details, see Andrews (1954). All the ARE results stated for the median test apply equally to the control median test, since these two tests have an ARE of 1, regardless of the parent distribution.

Finally, we discuss the chi-square test for k proportions as an approximate test for identical populations when the populations are all Bernoulli and the only relevant parameters are the probabilities of success.

PROBLEMS

10.1. Generate by enumeration the exact null probability distribution of the k-sample median test statistic for $k = 3, n_1 = 2, n_2 = 1, n_3 = 1$. If the rejection region consists of those arrangements which are least likely under the null hypothesis, find this region R and the exact α. Compute the values of the Q statistic for all arrangements and compare that critical region for the same value of α with the region R.

10.2. Generate the exact distribution of the Kruskal-Wallis statistic H for the same k and n_i as in Problem 10.1. Find the critical region which consists of those rank sums R_1, R_2, R_3 which have the largest value of H and find exact α.

10.3. By enumeration, place the median test criterion (U_1, U_2, U_3) and the H test criterion (R_1, R_2, R_3) in one-to-one correspondence for the same k and n_i as in Problems 10.1 and 10.2. If the two tests reject for the largest values of Q and H, respectively, which test seems to distinguish better between extreme arrangements?

10.4. Verify that the form of H given in (4.4) is algebraically equivalent to (4.2).

10.5. Show that H with $k = 2$ is exactly equivalent to the large-sample approximation to the two-sample Wilcoxon rank-sum test statistic mentioned in Section 8.2 with $m = n_1, n = n_2$.

10.6. Show that H is equivalent to the F test statistic in one-way analysis-of-variance problem if applied to the ranks of the observations rather than the actual numbers.

Hint: Express the F ratio as a function of H in the form given in (4.3) or (4.2) to show that

$$F = \left[\frac{k-1}{N-k}\left(\frac{N-1}{H} - 1\right)\right]^{-1}$$

This is an example of what is called a rank transform statistic. For related interesting results, see for example, Iman and Conover (1981).

10.7. Write the k-sample median test statistic given in (2.2) in the form of (5.1) (cf. Problem 7.2).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

10.8. How could the subsampling procedure described in Section 9.9 be extended to test the equality of variances in k populations?

10.9. In the context of the k-sample control median test defined in Section 10.3, show that for any $i(=2,3,\ldots,k)$ and $j(=0,1,\ldots,q)$ the random variable V_{ij} has a binomial distribution. What are the parameters of the distribution (i) in general and (ii) under H_0?

10.10. Show that under H_0: $\theta_1=\theta_2=\cdots=\theta_k$, the distribution of V^*, the sum test statistic for comparisons with a control when θ_1 is known in the case of unequal sample sizes defined in Section 10.7, is given by

$$p[V^*=v]=\prod_{1=2}^{k}\binom{n_i}{v_i}(0.5)^N$$

where $N=\sum_{i=2}^{k}n_i, v_i=0,1,\ldots,n_i, i=1,2,\ldots,k$, are such that $\sum_{1=2}^{k=v}(v_i/n_i)=v$. Enumerate the probability distribution of V^* when (i) $n_1=n_2=n_3=3$ and (ii) $n_1=3, n_2=n_3=2$.

10.11. In the context of the Jonckheere-Terpstra test discussed in Section 10.6, show that under H_0 for $i<j<r$,

$$\mathrm{cov}(U_{ij},U_{ir}\,|\,H_0)=\mathrm{cov}(U_{ji},U_{ri}\,|\,H_0)=n_in_jn_r/12$$

$$\mathrm{cov}(U_{ij},U_{ri}\,|\,H_0)=\mathrm{cov}(U_{ji},U_{ir}\,|\,H_0)=-n_in_jn_r/12$$

Hint: $2\ \mathrm{cov}(U_{ij},U_{ir})=\mathrm{var}(U_{ij},U_{ir})-\mathrm{var}(U_{ij})+\mathrm{var}(U_{ir})=\mathrm{var}(U_{i,j+r})-\mathrm{var}(U_{ij})+\mathrm{var}(U_{ir})$, where $U_{i,j+r}$ is the Mann-Whitney U statistic computed between the ith sample and the combined jth and rth samples.

10.12. Struckman-Johnson (1988) surveyed 623 students in a study to compare the proportions of men and women at a small midwestern university who have been coerced by their date into having sexual intercourse (date rape). A survey of over 600 students produced 623 responses. Of the 355 female respondents, 79 reported an experience of coercion, while 43 of the 268 male respondents reported coercion. Test the null hypothesis that males and females experience coercion at an equal rate.

10.13. Many psychologists have developed theories about how different kinds of brain dominance may affect recall ability of information presented in various formats. Brown and Evans (1986) compared recall ability of subjects classified into three groups according to their approach to problem solving as a result of their scores on the Human Information Process Survey. The three groups are Left (active, verbal, logical), Right (receptive, spacial, intuitive), and Integrative (combination of right and left). Information was presented to these subjects in tabular form about the number of physicians who practice in six different states. Recall was measured by how accurately the subjects were able to rank the states from highest to lowest after the presentation concluded. For the scores in Table 1 determine whether median recall ability is the same for the three groups (higher scores indicate greater recall).

Table 1

Left	*Right*	*Integrative*
35	17	28
32	20	30
38	25	31
29	15	25
36	10	26
31	12	24
33	8	24
35	16	27

10.14. A matching-to-sample (MTS) task is used by psychologists to understand how other species perceive and use identity relations. A standard MTS task consists of having subjects observe a sample stimulus and then rewarding the subject if it responds to an identical (matching) sample stimulus. Then the psychologist studies the ability of subjects to transfer the matching concept to other sample stimuli. Oden, Thompson, and Premack (1988) reported a study in which four infant chimpanzees learned an MTS task with only two training sample stimuli. Then the chimpanzees were tested on their ability to transfer the learning to three kinds of novel items, classified as Objects, Fabrics, and Food. The data were recorded as number of correct matches in a total of 24 trails. One purpose of the study was to show that the concept of matching is broadly construed by chimpanzees irrespective of the type of sample stimulus. Determine whether the data in Table 2 support this theory.

Table 2

Chimp	*Training*	*Object*	*Fabric*	*Food*
Whiskey	20	22	22	18
Liza	23	19	22	13
Opal	18	20	18	15
Frieda	21	21	19	19

10.15. Andrews (1989) examines attitudes toward advertising by undergraduate marketing students at universities in six different geographic regions. Attitudes were measured by answers to a questionnaire that allowed responses on a 7-point Likert scale ($1=$strongly disagree and $7=$strongly agree). Three statements on the questionnaire related to the social dimension were (1) most advertising insults the intelligence of the average consumer; (2) advertising often persuades people to buy things they shouldn't buy; and (3) in general, advertisements present a true picture of the product being advertised. For the mean scores given in Table 3, determine whether there are any regional differences in attitude for the social dimension.

Table 3

Region	*Insults*	*Persuades*	*True picture*
Northwest	3.69	4.48	3.69
Midwest	4.22	3.75	3.25
Northeast	3.63	4.54	4.09
Southwest	4.16	4.35	3.61
South Central	3.96	4.73	3.41
Southeast	3.78	4.49	3.64

10.16. Prior to the Alabama-Auburn football game, 80 Alabama alumni, 75 Auburn alumni, and 45 residents of Tuscaloosa who are not alumni of either are asked who they think will win the game. The responses are as follows:

	Alabama	*Auburn*	*Tuscaloosa*
Alabama win	55	15	30
Auburn win	25	60	15

Do the three groups have the same probability of thinking Alabama will win?

10.17. Random samples of 100 insurance company executives, 100 transportation company executives, and 100 media company executives were classified according to highest level of formal education using the code 10 = some college, 20 = bachelor's degree, 30 = master's degree, 40 = more than master's. The results are shown below. Determine whether median education level is the same for the three groups at $\alpha = 0.05$.

Education	*Insurance*	*Transportation*	*Media*
10	19	31	33
20	20	37	34
30	36	20	21
40	25	12	12
Total	100	100	100

10.18. Four different experimental methods of treating schizophrenia—(1) weekly shock treatments, (2) weekly treatments of carbon dioxide inhalations, (3) biweekly shock treatment alternated with biweekly carbon dioxide inhalations, and (4) tranquilizer drug treatment—are compared by assigning a group of schizophrenic patients randomly into four treatment groups. The data below are the number of patients who did

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and did not improve in four weeks of treatment. Test the null hypothesis that the treatments are equally effective.

Treatment	*Number improved*	*Number not improved*
1	43	12
2	24	28
3	32	16
4	29	24

10.19. Eighteen fish of a comparable size in a particular variety are divided randomly into three groups and each group is prepared by a different chef using the same recipe. Each prepared fish is then rated on each of the criteria of aroma, flavor, texture, and moisture by professional tasters. Use the composite scores below to test the null hypothesis that mean scores for all three chefs are the same.

Chef A	*Chef B*	*Chef C*
4.05	4.35	2.24
5.04	3.88	3.93
3.45	3.02	3.37
3.57	4.56	3.21
4.23	4.37	2.35
4.18	3.31	2.59

10.20. An office has three computers, A, B, and C. In a study of computer usage, the firm has kept records on weekly use rates for 7 weeks, except that computer A was out for repairs for part of 2 weeks. The eventual goal is to decide which computers to put under a service contract because they have a higher usage rate. As a first step in this study, analyze the data below on weekly computer usage rates to determine whether there is a significant difference in average usage. Can you make a preliminary recommendation?

A	*B*	*C*
12.3	15.7	32.4
15.4	10.8	41.2
10.3	45.0	35.1
8.0	12.3	25.0
14.6	8.2	8.2
	20.1	18.4
	26.3	32.5

10.21. A company is testing four cereals to determine taste preferences of potential buyers. Four different panels of persons are selected independently; one cereal is

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

presented to all members of each panel. After testing, each person is asked if he would purchase the product. The results are shown below. Test the null hypothesis that taste preference is the same for each cereal.

	Cereal			
	A	*B*	*C*	*D*
Number who would buy	75	80	57	80
Number who would not buy	50	60	43	70

10.22. Below are four sets of five measurements, each set an array of data (for your convenience) on the smoothness of a certain type of paper, each set obtained from a different laboratory. Find an approximate P value to test whether the median smoothness can be regarded as the same for all laboratories.

Laboratory			*Data*		
A	38.7	41.5	43.8	44.5	45.5
B	39.2	39.3	39.7	41.4	41.8
C	34.0	35.0	39.0	40.0	43.0
D	34.1	34.8	34.9	35.4	37.2

10.23. Verify the value of the Kruskal-Wallis test statistic given in the SAS solution to Example 2.1 in Chapter 8.

11

Measures of Association for Bivariate Samples

11.1 INTRODUCTION: DEFINITION OF MEASURES OF ASSOCIATION IN A BIVARIATE POPULATION

In Chapter 5 we saw that the ordinary sign test and the Wilcoxon signed-rank test procedures, although discussed in terms of inferences in a single-sample problem, could be applied to paired-sample data by basing the statistical analysis on the differences between the pairs of observations. The inferences then must be concerned with the population of differences as opposed to some general relationship between the two dependent random variables. One parameter of this population of differences, the variance, does contain information concerning their relationship, since

$$\mathrm{var}(X - Y) = \mathrm{var}(X) + \mathrm{var}(Y) - 2\,\mathrm{cov}(X, Y)$$

It is this covariance factor and a similar measure with which we shall be concerned in this chapter.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

In general, if X and Y are two random variables with a bivariate probability distribution, their covariance, in a certain sense, reflects the direction and amount of association or correspondence between the variables. The covariance is large and positive if there is a high probability that large (small) values of X are associated with large (small) values of Y. On the other hand, if the correspondence is inverse so that large (small) values of X generally occur in conjunction with small (large) values of Y, their covariance is large and negative. This comparative type of association is referred to as *concordance* or *agreement*. The covariance parameter as a measure of association is difficult to interpret because its value depends on the orders of magnitude and units of the random variables concerned. A nonabsolute or relative measure of association circumvents this difficulty. The Pearson product-moment correlation coefficient, defined as

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{[\operatorname{var}(X)\ \operatorname{var}(Y)]^{1/2}}$$

is a measure of the linear relationship between X and Y. This coefficient is invariant under changes of scale and location in X and Y, and in classical statistics this parameter is usually used as the relative measure of association in a bivariate distribution. The absolute value of the correlation coefficient does not exceed 1, and its sign is determined by the sign of the covariance. If X and Y are independent random variables, their correlation is zero, and therefore the magnitude of ρ in some sense measures the degree of association. Although it is not true in general that a zero correlation implies independence, the bivariate normal distribution is a significant exception, and therefore in the normal-theory model ρ is a good measure of association. For random variables from other bivariate populations, ρ may not be such a good description of relationship since dependence may be reflected in a wide variety of types of relationships. One can only say in general that ρ is a more descriptive measure of dependence than covariance because ρ does not depend on the scales of X and Y.

If the main justification for the use of ρ as a measure of association is that the bivariate normal is such an important distribution in classical statistics and zero correlation is equivalent to independence for that particular population, this reasoning has little significance in nonparametric statistics. Other population measures of association should be equally acceptable, but the approach to measuring relationships might be analogous, so that interpretations are simplified. Because ρ is so widely known and accepted, any other measure would preferably emulate its properties.

Suppose we define a "good" relative measure of association as one which satisfies the following criteria:

1. For any two independent pairs (X_i, Y_i) and (X_j, Y_j) of random variables which follow this bivariate distribution, the measure will equal $+1$ if the relationship is direct and perfect in the sense that

   $$X_i < X_j \text{ whenever } Y_i < Y_j \qquad \text{or} \qquad X_i > X_j \text{ whenever } Y_i > Y_j$$

   This relation will be referred to as *perfect concordance* (agreement).
2. For any two independent pairs, the measure will equal -1 if the relationship is indirect and perfect in the sense that

   $$X_i < X_j \text{ whenever } Y_i > Y_j \qquad \text{or} \qquad X_i > X_j \text{ whenever } Y_i < Y_j$$

   This relation will be referred to as *perfect discordance* (disagreement).
3. If neither criterion 1 nor criterion 2 is true for all pairs, the measure will lie between the two extremes -1 and $+1$. It is also desirable that, in some sense, increasing *degrees of concordance* are reflected by increasing positive values, and increasing *degrees of discordance* are reflected by increasing negative values.
4. The measure will equal zero if X and Y are independent.
5. The measure for X and Y will be the same as for Y and X, or $-X$ and $-Y$, or $-Y$ and $-X$.
6. The measure for $-X$ and Y or X and $-Y$ will be the negative of the measure for X and Y.
7. The measure will be invariant under all transformations of X and Y for which order of magnitude is preserved.

The parameter ρ is well known to satisfy the first six of these criteria. It is a type of measure of concordance in the same sense that covariance measures the degree to which the two variables are associated in magnitude. However, although ρ is invariant under positive linear transformations of the random variables, it is not invariant under all order-preserving transformations. This last criterion seems especially desirable in nonparametric statistics, as we have seen that in order to be distribution-free, inferences must usually be determined by relative magnitudes as opposed to absolute magnitudes of the variables under study. Since probabilities of events involving only inequality relations between random variables are invariant under all order-preserving transformations, a measure of association which is a function of the probabilities of concordance and discordance will satisfy the seventh criterion. Perfect direct and indirect association between X and Y are reflected by perfect concordance and perfect discordance, respectively, and in the same spirit as ρ measures a perfect direct and indirect linear relationship between the variables. Thus an appropriate combination of these probabilities will provide a measure of association which will satisfy all seven of these desirable criteria.

For any two independent pairs of random variables (X_i, Y_i) and (X_j, Y_j), we denote by p_c and p_d the probabilities of concordance and discordance, respectively.

$$\begin{aligned} p_c &= P\{[(X_i < X_j) \cap (Y_i < Y_j)] \cup [(X_i > X_j) \cap (Y_i > Y_j)] \\ &= P[(X_j - X_i)(Y_j - Y_i) > 0] \\ &= P[(X_i < X_j) \cap (Y_i < Y_j)] + P[(X_i > X_j) \cap (Y_i > Y_j)] \end{aligned}$$

$$\begin{aligned} p_d &= P[(X_j - X_i)(Y_j - Y_i) < 0] \\ &= P[(X_i < X_j) \cap (Y_i > Y_j)] + P[(X_i > X_j) \cap (Y_i < Y_j)] \end{aligned}$$

Perfect association between X and Y is reflected by either perfect concordance or perfect discordance, and thus some combination of these probabilities should provide a measure of association. The *Kendall coefficient* τ is defined as the difference

$$\tau = p_c - p_d$$

and this measure of association satisfies our desirable criteria 1 to 7. If the marginal probability distributions of X and Y are continuous, so that the probability of ties $X_i = X_j$ or $Y_i = Y_j$ within groups is eliminated, we have

$$\begin{aligned} p_c &= \{P(Y_i < Y_j) - P[(X_i > X_j) \cap (Y_i < Y_j)] \\ &\quad + \{P(Y_i > Y_j) - P[(X_i < X_j) \cap (Y_i > Y_j)]\} \\ &= P(Y_i < Y_j) + P(Y_i > Y_j) - p_d \\ &= 1 - p_d \end{aligned}$$

Thus in this case τ can also be expressed as

$$\tau = 2p_c - 1 = 1 - 2p_d$$

How does τ measure independence? If X and Y are independent and continuous random variables, $P(X_i < X_j) = P(X_i > X_j)$ and further

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

the joint probabilities in p_c or p_d are the product of the individual probabilities. Using these relations, we can write

$$\begin{aligned} p_c &= P(X_i < X_j)P(Y_i < Y_j) + P(X_i > X_j)P(Y_i > Y_j) \\ &= P(X_i > X_j)P(Y_i < Y_j) + P(X_i < X_j)P(Y_i > Y_j) = p_d \end{aligned}$$

and thus $\tau = 0$ for independent continuous random variables. In general, the converse is not true, but this disadvantage is shared by ρ. For the bivariate normal population, however, $\tau = 0$ if and only if $\rho = 0$, that is, if and only if X and Y are independent. This fact follows from the relation

$$\tau = \frac{2}{\pi} \arcsin \rho$$

which can be derived as follows. Suppose that X and Y are bivariate normal with variances σ_X^2 and σ_Y^2 and correlation coefficient ρ. Then for any two independent pairs (X_i, Y_i) and (X_j, Y_j) from this population, the differences

$$U = \frac{X_i - X_j}{\sqrt{2}\sigma_X} \qquad \text{and} \qquad V = \frac{Y_i - Y_j}{\sqrt{2}\sigma_Y}$$

also have a bivariate normal distribution, with zero means, unit variances, and covariance equal to ρ. Thus $\rho(U, V) = \rho(X, Y)$. Since

$$p_c = P(UV > 0)$$

we have

$$\begin{aligned} p_c &= \int_{-\infty}^{0} \int_{-\infty}^{0} \varphi(x, y)\, dx\, dy + \int_{0}^{\infty} \int_{0}^{\infty} \varphi(x, y)\, dx\, dy \\ &= 2 \int_{-\infty}^{0} \int_{-\infty}^{0} \varphi(x, y)\, dx\, dy = 2\Phi(0, 0) \end{aligned}$$

where $\varphi(x, y)$ and $\Phi(x, y)$ denote the density and cumulative distributions, respectively, of a standardized bivariate normal probability distribution. Since it can be shown that

$$\Phi(0, 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin \rho$$

we see that for the bivariate normal

$$p_c = \frac{1}{2} + \frac{1}{\pi} \arcsin \rho$$

and

$$\tau = \frac{2}{\pi} \arcsin \rho$$

In this chapter, the problem of point estimation of these two population measures of association, ρ and τ, will be considered. We shall find estimates which are distribution-free and discuss their individual properties and procedures for hypothesis testing, and the relationship between the two estimates will be determined. Another measure of association will be discussed briefly.

11.2 KENDALL'S TAU COEFFICIENT

In Section 11.1, Kendall's tau, a measure of association between random variables from any bivariate population, was defined as

$$\tau = p_c - p_d \tag{2.1}$$

where, for any two independent pairs of observations $(X_i, Y_i), (X_j, Y_j)$ from the population,

$$p_c = P[(X_j - X_i)(Y_j - Y_i) > 0] \quad \text{and} \quad p_d = P[(X_j - X_i)(Y_j - Y_i) < 0] \tag{2.2}$$

In order to estimate the parameter τ from a random sample of n pairs

$$(X_1, Y_1),\ (X_2, Y_2), \ldots, (X_n, Y_n)$$

drawn from this bivariate population, we must find point estimates of the probabilities p_c and p_d. For each set of pairs (X_i,Y_i), (X_j,Y_j) of sample observations, define the indicator variables

$$A_{ij} = \operatorname{sgn}(X_j - X_i)\ \operatorname{sgn}(Y_j - Y_i) \tag{2.3}$$

where

$$\operatorname{sgn}(u) = \begin{cases} -1 & \text{if } u < 0 \\ 0 & \text{if } u = 0 \\ 1 & \text{if } u > 0 \end{cases}$$

Then the values assumed by A_{ij} are

$$a_{ij} = \begin{cases} 1 & \text{if these pairs are concordant} \\ -1 & \text{if these pairs are discordant} \\ 0 & \text{if these pairs are neither concordant nor} \\ & \quad \text{discordant because of a tie in either component} \end{cases}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The marginal probability distribution of these indicator variables is

$$f_{A_{ij}}(a_{ij}) = \begin{cases} p_c & \text{if } a_{ij} = 1 \\ p_d & \text{if } a_{ij} = -1 \\ 1 - p_c - p_d & \text{if } a_{ij} = 0 \end{cases} \tag{2.4}$$

and the expected value is

$$E(A_{ij}) = 1p_c + (-1)p_d = p_c - p_d = \tau \tag{2.5}$$

Since obviously we have $a_{ij} = a_{ji}$ and $a_{ii} = 0$, there are only $\binom{n}{2}$ sets of pairs which need be considered. An unbiased estimator of τ is therefore provided by

$$T = \sum\sum_{1 \leqslant i<j \leqslant n} \frac{A_{ij}}{\binom{n}{2}} = 2 \sum\sum_{1 \leqslant i<j \leqslant n} \frac{A_{ij}}{n(n-1)} \tag{2.6}$$

This measure of the association in the paired-sample observations is called *Kendall's sample tau coefficient*.

The reader should note that with the definition of A_{ij} in (2.3) that allows for tied observations, no assumption regarding the continuity of the population was necessary, and thus T is an unbiased estimator of the parameter τ in *any* bivariate distribution. Since the variance of T approaches zero as the sample size approaches infinity, T is also a consistent estimator of τ for any bivariate distribution, as we now show.

In order to determine the variance of T, the variances and covariances of the A_{ij} must be evaluated since T is a linear combination of these indicator random variables. From (2.6), we have

$$n^2(n-1)^2\operatorname{var}(T) = 4\left[\sum\sum_{1 \leqslant i<j \leqslant n} \operatorname{var}(A_{ij}) + \sum\sum\sum\sum_{\substack{1 \leqslant i<j \leqslant n \\ 1 \leqslant h<k \leqslant n \\ i \neq h \text{ or } j \neq k}} \operatorname{cov}(A_{ij}, A_{hk})\right] \tag{2.7}$$

Since the A_{ij} are identically distributed for all $i < j$, and A_{ij} and A_{hk} are independent for all $i \neq h$ and $j \neq k$ (no pairs in common), (2.7) can be written as

$$n^2(n-1)^2\,\mathrm{var}(T) = 4\Bigg[\binom{n}{2}\underset{i<j}{\mathrm{var}}(A_{ij}) + \sum_{i=1}^{n-1}\sum_{\substack{j=i+1\\ j\neq k}}^{n}\sum_{k=i+1}^{n}\mathrm{cov}(A_{ij},A_{ik})$$
$$+\sum_{j=2}^{n}\sum_{\substack{i=1\\ i\neq k}}^{j-1}\sum_{k=1}^{j-1}\mathrm{cov}(A_{ij},A_{kj})$$
$$+\sum_{j=2}^{n}\sum_{\substack{i=1\\ i\neq k}}^{j-1}\sum_{k=j+1}^{n}\mathrm{cov}(A_{ij},A_{jk})$$
$$+\sum_{i=2}^{n-1}\sum_{\substack{j=i+1\\ j\neq k}}^{n}\sum_{k=1}^{i-1}\mathrm{cov}(A_{ij},A_{ki})\Bigg] \tag{2.8}$$

By symmetry, all of the covariance terms in (2.8) are equal. They are grouped together according to which of the (X,Y) pairs are common to the (A_{ij}, A_{hk}) in order to facilitate counting the number of terms in each summation set. Within the first set we have two distinct permutations, (A_{ij},A_{ik}) and (A_{ik},A_{ij}), for each of the $\binom{n}{2}$ choices of $i \neq j \neq k$, and similarly for the second set. But the third and fourth sets do not allow for reversal of the A_{ij} and A_{hk} terms since this makes a different (X,Y) pair in common, and so there are only $\binom{n}{3}$ covariance terms in each of these summations. The total number of distinguishable covariance terms then is $(2+2+1+1)\binom{n}{3} = 6\binom{n}{3}$, and (2.8) can be written as simply

$$n^2(n-1)^2\ \mathrm{var}(T) = 4\left[\binom{n}{2}\mathrm{var}(A_{ij}) + 6\binom{n}{3}\mathrm{cov}(A_{ij},A_{ik})\right]$$

or

$$n(n-1)\mathrm{var}(T) = 2\ \mathrm{var}(A_{ij}) + 4(n-2)\mathrm{cov}(A_{ij},A_{ik}) \tag{2.9}$$

for any

$$i<j;\ i<k;\ j\neq k;\ i=1,2,\ldots,n-1;\ j=2,3,\ldots,n;\ k=2,3,\ldots,n$$

Using the marginal probability distribution of A_{ij} given in (2.4), the variance of A_{ij} is easily evaluated as follows:

$$\begin{aligned} E(A_{ij}^2) &= 1p_c + (-1)^2 p_d = p_c + p_d \\ \mathrm{var}(A_{ij}) &= (p_c + p_d) - (p_c - p_d)^2 \end{aligned} \tag{2.10}$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

The covariance expression, however, requires knowledge of the joint distribution of A_{ij} and A_{ik}, which can be expressed as

$$f_{A_{ij},A_{ik}}(a_{ij},a_{ik})=\begin{cases} p_{cc} & \text{if } a_{ij}=a_{ik}=1 \\ p_{dd} & \text{if } a_{ij}=a_{ik}=-1 \\ p_{cd} & \text{if } a_{ij}=1,\ a_{ik}=-1 \\ & \quad \text{or } a_{ij}=-1,\ a_{ik}=1 \\ 1-p_{cc}-p_{dd}-2p_{cd} & \text{if } a_{ij}=0,\ a_{ik}=-1,0,1 \\ & \quad \text{or } a_{ij}=-1,0,1,\ a_{ik}=0 \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

for all $i<j$, $i<k$, $j \neq k$, $i=1,2,\ldots,n$, and some $0 \leqslant p_{cc}, p_{dd}, p_{cd} \leqslant 1$. Thus we can evaluate

$$\begin{aligned} E(A_{ij}A_{ik}) &= 1^2 p_{cc} + (-1)^2 p_{dd} + 2(-1)p_{cd} \\ \operatorname{cov}(A_{ij},A_{ik}) &= p_{cc}+p_{dd}-2p_{cd}-(p_c-p_d)^2 \end{aligned} \tag{2.12}$$

Substitution of (2.10) and (2.12) in (2.9) gives

$$\begin{aligned} n(n-1)\operatorname{var}(T) &= 2(p_c+p_d)+4(n-2)(p_{cc}+p_{dd}-2p_{cd}) \\ &\quad -2(2n-3)(p_c-p_d)^2 \end{aligned} \tag{2.13}$$

so that the variance of T is of order $1/n$ and therefore approaches zero as n approaches infinity.

The results obtained so far are completely general, applying to all random variables. If the marginal distributions of X and Y are continuous, $P(A_{ij}=0)=0$ and the resulting identities

$$p_c+p_d=1 \qquad \text{and} \qquad p_{cc}+p_{dd}+2p_{cd}=1$$

allow us to simplify (2.13) to a function of, say, p_c and p_{cd} only:

$$\begin{aligned} n(n-1)\ \operatorname{var}(T) &= 2-2(2n-3)(2p_c-1)^2+4(n-2)(1-4p_{cd}) \\ &= 8(2n-3)p_c(1-p_c)-16(n-2)p_{cd} \end{aligned} \tag{2.14}$$

Since for X and Y continuous we also have

$$\begin{aligned} p_{cd} &= P(A_{ij}=1 \cap A_{ik}=-1) \\ &= P(A_{ij}=1)-P(A_{ij}=1 \cap A_{ik}=1) \\ &= p_c - p_{cc} \end{aligned}$$

another expression equivalent to (2.14) is

$$
\begin{aligned}
n(n-1)\operatorname{var}(T) &= 8(2n-3)p_c(1-p_c) - 16(n-2)(p_c - p_{cc}) \\
&= 8p_c(1-p_c) + 16(n-2)(p_{cc} - p_c^2)
\end{aligned}
\tag{2.15}
$$

We have already interpreted p_c as the probability that the pair (X_i,Y_i) is concordant with (X_j,Y_j). Since the parameter p_{cc} is

$$
\begin{aligned}
p_{cc} &= P(A_{ij} = 1 \cap A_{ik} = 1) \\
&= P[(X_j - X_i)(Y_j - Y_i) > 0 \cap (X_k - X_i)(Y_k - Y_i) > 0]
\end{aligned}
\tag{2.16}
$$

for all $i < j$, $i < k$, $j \neq k$, $i = 1, 2, \ldots, n$, we interpret p_{cc} as the probability that the pair (X_i,Y_i) is concordant with both (X_j,Y_j) the (X_k,Y_k).

Integral expressions can be obtained as follows for the probabilities p_c and p_{cc} for random variables X and Y from any continuous bivariate population $F_{X,Y}(x,y)$.

$$
\begin{aligned}
p_c &= P[(X_i < X_j) \cap (Y_i < Y_j)] + P[(X_i > X_j) \cap (Y_i > Y_j)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P[(X_i < x_j) \cap (Y_i < y_j)] f_{X_i,Y_i}(x_j, y_j)\, dx_j\, dy_j \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P[(X_j < x_i) \cap (Y_j < y_i)] f_{X_i,Y_i}(x_i, y_i)\, dx_i\, dy_i \\
&= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{X,Y}(x,y) f_{X,Y}(x,y)\, dx\, dy
\end{aligned}
\tag{2.17}
$$

$$
\begin{aligned}
p_{cc} &= P(\{[(X_i < X_j) \cap (Y_i < Y_j)] \cup [(X_i > X_j) \cap (Y_i > Y_j)]\} \\
&\quad \cap \{[(X_i < X_k) \cap (Y_i < Y_k) \cup [(X_i > X_k) \cap (Y_i > Y_k)]\}) \\
&= P[(A \cup B) \cap (C \cup D)] \\
&= P[(A \cap C) \cup (B \cap D) \cup (A \cap D) \cup (B \cap C)] \\
&= P(A \cap C) + P(B \cap D) + 2P(A \cap D) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{P[(X_j > x_i) \cap (Y_j > y_i) \cap (X_k > x_i) \cap (Y_k > y_i)] \\
&\qquad + P[(X_j < x_i) \cap (Y_j < y_i) \cap (X_k < x_i) \cap (Y_k < y_i)] \\
&\qquad + 2P[(X_j > x_i) \cap (Y_j > y_i) \cap (X_k < x_i) \cap (Y_k < y_i)]\} \\
&\qquad \times f_{X_i,Y_i}(x_i, y_i)\, dx_i\, dy_i \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\{P[(X > x) \cap (Y > y)]\}^2 + \{P[(X < x) \cap (Y < y)]\}^2 \\
&\qquad + 2P[(X > x) \cap (Y > y)] P[(X < x) \cap (Y < y)]) f_{X,Y}(x,y)\, dx\, dy
\end{aligned}
$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{P[(X > x) \cap (Y > y)] + P[(X < x) \cap (Y < y)]\}^2 f_{X,Y}(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [1 - F_X(x) - F_Y(y) + 2F_{X,Y}(x,y)]^2 f_{X,Y}(x,y)\,dx\,dy \qquad (2.18)$$

Although T as given in (2.6) is perhaps the simplest form for deriving theoretical properties, the coefficient can be written in a number of other ways. In terms of all n^2 pairs for which A_{ij} is defined, (2.6) can be written as

$$T = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{A_{ij}}{n(n-1)} \qquad (2.19)$$

Now we introduce the notation

$$U_{ij} = \operatorname{sgn}(X_j - X_i) \qquad \text{and} \qquad V_{ij} = \operatorname{sgn}(Y_j - Y_i)$$

so that $A_{ij} = U_{ij}V_{ij}$ for all i, j. Assuming that $X_i \neq X_j$ and $Y_i \neq Y_j$ for all $i \neq j$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} V_{ij}^2 = n(n-1)$$

and (2.19) can be written in a form resembling an ordinary sample correlation coefficient as

$$T = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}V_{ij}}{\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} V_{ij}^2\right)\right]^{1/2}} \qquad (2.20)$$

Kendall and Gibbons (1990) often use T in still another form, which arises by simply classifying sets of differences according to the resulting sign of A_{ij}. If C and Q denote the number of positive and negative A_{ij} for $1 \leqslant i < j \leqslant n$, respectively, and the total is $S = C - Q$, we have

$$T = (C - Q)\Big/\binom{n}{2} = S\Big/\binom{n}{2} \qquad (2.21)$$

If there are no ties within either the X or Y groups, that is, $A_{ij} \neq 0$ for $i \neq j$, $C + Q = \binom{n}{2}$ and (2.21) can be written as

$$T = \frac{2C}{\binom{n}{2}} - 1 = 1 - \frac{2Q}{\binom{n}{2}} \tag{2.22}$$

These two forms in (2.22) are analogous to the expression in Section 1 for the parameter

$$\tau = 2p_c - 1 = 1 - 2p_d$$

and $C/\binom{n}{2}$ and $Q/\binom{n}{2}$ are obviously unbiased estimators for p_c and p_d, respectively. The quantity C is perhaps the simplest to calculate for a given sample of n pairs. Assuming that the pairs are written from smallest to largest according to the value of the X component, C is simply the number of values of $1 \leqslant i < j \leqslant n$ for which $Y_j - Y_i > 0$, since only then shall we have $a_{ij} = 1$.

Another interpretation of T is as a *coefficient of disarray*, since it can be shown (see Kendall and Gibbons, 1990, pp. 30–31) that the total number of interchanges between two consecutive Y observations required to transform the Y arrangement into the natural ordering from smallest to largest, i.e., to transform the Y arrangement into the X arrangement, is equal to Q, or $\binom{n}{2}(1-T)/2$. This will be illustrated later in Section 11.6.

NULL DISTRIBUTION OF *T*

Suppose we wish to test the null hypothesis that the X and Y random variables are independent. Since $\tau = 0$ for independent variables, the null distribution of T is symmetric about the origin. For a general alternative of nonindependence, the rejection region of size α then should be

$$T \in R \qquad \text{for } |T| \geqslant t_{\alpha/2}$$

where $t_{\alpha/2}$ is chosen so that

$$P(|T| \geqslant t_{\alpha/2} | H_0) = \alpha$$

For an alternative of positive dependence, a similar one-sided critical region is appropriate.

We must now determine the random sampling distribution of T under the assumption of independence. For this purpose, it will be more convenient, but not necessary, to assume that the X and Y sample observations have both been ordered from smallest to largest and assigned positive integer ranks. The data then consist of n sets of

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

pairs of ranks. The justification for this assumption is that, like τ, T is invariant under all order-preserving transformations. Its numerical value then depends only on the relative magnitudes of the observations and is the same whether calculated for variate values or ranks. For samples with no ties, the $n!$ distinguishable pairings of ranks are all equally likely under the null hypothesis. The value of T is completely determined by the value of C or S because of the expressions in (2.21) and (2.22), and it is more convenient to work with C. Denote by $u(n, c)$ the number of pairings of n ranks which result in exactly c positive a_{ij}, $1 \leqslant i < j \leqslant n$. Then

$$P(C = c) = \frac{u(n, c)}{n!} \tag{2.23}$$

and

$$f_T(t) = P(T = t) = P\left[C = \binom{n}{2}\frac{t+1}{2}\right] \tag{2.24}$$

We shall now find a *recursive relation* to generate the values of $u(n+1, c)$ from knowledge of the values of $u(n, c)$ for some n and all c. Assuming that the observations are written in order of magnitude of the X component, the value of C depends only on the resulting permutation of the Y ranks. If s_i denotes the rank of the Y observation which is paired with the rank i in the X sample, for $i = 1, 2, \ldots, n, c$ equals the number of integers greater than s_1, plus the number of integers greater than s_2 excluding s_1, plus the number exceeding s_3 excluding s_1 and s_2, etc. For any given permutation of n integers which has this sum c, we need only consider what insertion of the number $n+1$ in any of the $n+1$ possible positions of the permutation $(s_1, s_2, \ldots, s_n)$ does to the value of c. If $n+1$ is in the first position, c is clearly unchanged. If $n+1$ is in the second position, there is one additional integer greater than s_1, so that c is increased by 1. If in the third position, there is one additional integer greater than both s_1 and s_2, so that c is increased by 2. In general, if $n+1$ is in the kth position, c is increased by $k-1$ for all $k = 1, 2, \ldots, n+1$. Therefore the desired recursive relation is

$$u(n+1, c) = u(n, c) + u(n, c-1) + u(n, c-2) + \cdots + u(n, c-n) \tag{2.25}$$

In terms of s, since for a set of n pairs

$$s = 2c - \frac{n(n-1)}{2} \tag{2.26}$$

insertion of $n+1$ in the kth position increases c by $k-1$, the new value s' of s for $n+1$ pairs will be

$$\begin{aligned} s' &= 2c' - \frac{n(n+1)}{2} = 2(c+k-1) - \frac{n(n+1)}{2} \\ &= 2c - \frac{n(n-1)}{2} + 2(k-1) - n = s + 2(k-1) - n \end{aligned}$$

In other words, s is increased by $2(k-1)-n$ for $k = 1, 2, \ldots, n+1$, and corresponding to (2.25) we have

$$\begin{aligned} u(n+1,\ s) = u(n,\ s+n) + u(n,\ s+n-2) + u(n, s+n-4) \\ + \cdots + u(n, s-n+2) + u(n, s-n) \end{aligned} \tag{2.27}$$

The distribution of S is symmetrical about zero, and from (2.26) it is clear that S for n pairs is an even or odd integer according as $n(n-1)/2$ is even or odd. Because of this symmetry, tables are most easily constructed for S (or T) rather than C or Q. The null distribution of S is given in Table L of the Appendix. More extensive tables of the null distribution of S or T are given in Kaarsemaker and Van Wijngaarden (1952, 1953), Best (1973, 1974), Best and Gipps (1974), Nijsse (1988), and Kendall and Gibbons (1990).

A simple example will suffice to illustrate the use of (2.25) or (2.27) to set up tables of these probability distributions. When $n = 3$, the 3! permutations of the Y ranks and the corresponding values of C and S are:

Permutation	*123*	*132*	*213*	*231*	*312*	*321*
c	3	2	2	1	1	0
s	3	1	1	−1	−1	−3

The frequencies then are:

c	0	1	2	3
s	−3	−1	1	3
$u(3,c)$ or $u(3,s)$	1	2	2	1

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For C, using (2.25), $u(4,c) = \sum_{i=0}^{3} u(3c - i)$, or

$$
\begin{aligned}
&u(4,0) = u(3,0) = 1\\
&u(4,1) = u(3,1) + u(3,0) = 3\\
&u(4,2) = u(3,2) + u(3,1) + u(3,0) = 5\\
&u(4,3) = u(3,3) + u(3,2) + u(3,1) + u(3,0) = 6\\
&u(4,4) = u(3,3) + u(3,2) + u(3,1) = 5\\
&u(4,5) = u(3,3) + u(3,2) = 3\\
&u(4,6) = u(3,3) = 1
\end{aligned}
$$

Alternatively, we could use (2.27), or $u(4,s) = \sum_{i=0}^{3} u(3, s + 3 - 2i)$. Therefore the probability distributions for $n = 4$ are:

c	0	1	2	3	4	5	6
s	−6	−4	−2	0	2	4	6
t	−1	−2/3	−1/3	0	1/3	2/3	1
$f(c, s, \text{or } t)$	1/24	3/24	5/24	6/24	5/24	3/24	1/24

The way in which the $u(n,\ s,\ \text{or}\ c)$ are built up by cumulative sums indicates that simple schemes for their generation may be easily worked out (see, for example, Kendall and Gibbons, 1990, pp. 91–92).

The exact null distribution is thus easily found for moderate n. Since T is a sum of random variables, it can be shown using general limit theorems for independent variables that the distribution of a standardized T approaches the standard normal distribution as n approaches infinity. To use this fact to facilitate inferences concerning independence in large samples, we need to determine the null mean and variance of T. Since T was defined to be an unbiased estimator of τ for any bivariate population and we showed in Section 1 that $\tau = 0$ for independent, continuous random variables, the mean is $E(T \mid H_0) = 0$. In order to find $\text{var}(T \mid H_0)$ for X and Y continuous, (2.15) is used with the appropriate p_c and p_{cc} under H_0. Under the assumption that X and Y have continuous marginal distributions and are independent, they can be assumed to be identically distributed according to the uniform distribution over the interval (0,1), because of the probability-integral transformation. Then, in (2.17) and (2.18), we have

$$
\begin{aligned}
p_c &= 2\int_0^1\int_0^1 xy\,dx\,dy = 1/2\\
p_{cc} &= \int_0^1\int_0^1 (1 - x - y + 2xy)^2\,dx\,dy = 5/18
\end{aligned}
\qquad (2.28)
$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Substituting these results in (2.15), we obtain

$$n(n-1)\ \mathrm{var}(T) = 2 + \frac{16(n-2)}{36}$$

$$\mathrm{var}(T) = \frac{2(2n+5)}{9n(n-1)} \tag{2.29}$$

For large n, the random variable

$$Z = \frac{3\sqrt{n(n-1)}\,T}{\sqrt{2(2n+5)}} \tag{2.30}$$

can be treated as a standard normal variable with density $\varphi(z)$.

If the null hypothesis of independence of X and Y is accepted, we can of course infer that the population parameter τ is zero. However, if the hypothesis is rejected, this implies dependence between the random variables but not necessarily that $\tau \neq 0$.

THE LARGE-SAMPLE NONNULL DISTRIBUTION OF KENDALL'S STATISTIC

The probability distribution of T is asymptotically normal for sample pairs from any bivariate population. Therefore, if any general mean and variance of T could be determined, T would be useful in large samples for other inferences relating to population characteristics besides independence. Since $E(T) = \tau$ for any distribution, T is particularly relevant in inferences concerning the value of τ. The expressions previously found for $\mathrm{var}(T)$ in (2.13) for any distribution and (2.15) for continuous distributions depend on unknown probabilities. Unless the hypothesis under consideration somehow determines $p_c,\ p_d,\ p_{cc},\ p_{dd}$, and p_{cd} (or simply p_c and p_{cc} for the continuous case), the exact variance cannot be found without some information about $f_{X,Y}(x,y)$. However, unbiased and consistent estimates of these probabilities can be found from the sample data to provide a consistent estimate $\hat{\sigma}(T)$ of the variance of T. The asymptotic distribution of $(T-\tau)/\hat{\sigma}(T)$ then remains standard normal.

Such estimates will be found here for paired samples containing no tied observations. We observed before that $C/\binom{n}{2}$ is an unbiased and consistent estimator of p_c. However, for the purpose of finding estimates for all the probabilities involved, it will be more convenient now to introduce a different notation. As before, we can assume without loss of generality that the n pairs are arranged in natural order according to the x component and that s_i is the rank of that y which is

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

paired with the ith smallest x for $i = 1, 2, \ldots, n$, so that the data are $(s_1, s_2, \ldots, s_n)$. Define

a_i = number of integers to the left of s_i and less than s_i

b_i = number of integers to the right of s_i and greater than s_i

Then

$c_i = a_i + b_i$ = number of values of $j = 1, 2, \ldots, n$ such that (x_i, y_i) is concordant with (x_j, y_j). There are $n(n-1)$ distinguishable sets of pairs, of which $\sum_{i=1}^{n} c_i$ are concordant. An unbiased estimate of p_c then is

$$\hat{p}_c = \sum_{i=1}^{n} \frac{c_i}{n(n-1)} \tag{2.31}$$

Similarly, we define

a'_i = number of integers to the left of s_i and greater than s_i

b'_i = number of integers to the right of s_i and less than s_i

and

$d_i = a'_i + b'_i$ = number of values of $j = 1, 2, \ldots, n$ such that (x_i, y_i) is discordant with (x_j, y_j). Then

$$\hat{p}_d = \sum_{i=1}^{n} \frac{d_i}{n(n-1)} \tag{2.32}$$

gives an unbiased estimate of p_d.

An unbiased and consistent estimate of p_{cc} is the number of sets of three pairs (x_i, y_i), (x_j, y_j), (x_k, y_k) for all $i \neq j \neq k$, for which the products $(x_i - x_j)(y_i - y_j)$ and $(x_i - x_k)(y_i - y_k)$ are both positive, divided by the number of distinguishable sets $n(n-1)(n-2)$. Denote by c_{ii} the number of values of j and k, $i \neq j \neq k, 1 \leqslant j, k \leqslant n$, such that (x_i, y_i) is concordant with both (x_j, y_j) and (x_k, y_k), so that

$$\hat{p}_{cc} = \sum_{i=1}^{n} \frac{c_{ii}}{n(n-1)(n-2)}$$

The pair (x_i, y_i) is concordant with both (x_j, y_j) and (x_k, y_k) if:

Group 1:	$s_j < s_i < s_k$	for $j < i < k$
	$s_k < s_i < s_j$	for $k < i < j$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Group 2:	$s_i < s_j < s_k$	for $i < j < k$
	$s_i < s_k < s_j$	for $i < k < j$
Group 3:	$s_j < s_k < s_i$	for $j < k < i$
	$s_k < s_j < s_i$	for $k < j < i$

Therefore c_{ii} is twice the sum of the following three corresponding numbers:

1. The number of unordered pairs of integers, one to the left and one to the right of s_i, such that the one to the left is less than s_i and the one to the right is greater than s_i.
2. The number of unordered pairs of integers, both to the right of s_i, such that both are greater than s_i.
3. The number of unordered pairs of integers, both to the left of s_i, such that both are less than s_i.

Then, employing the same notation as before, we have

$$c_{ii} = 2\left[\binom{a_i}{1}\binom{b_i}{1} + \binom{b_i}{2} + \binom{a_i}{2}\right] = (a_i + b_i)^2 - (a_i + b_i) = c_i^2 - c_i = c_i(c_i - 1)$$

and

$$\hat{p}_{cc} = \sum_{i=1}^{n} \frac{c_i(c_i - 1)}{n(n-1)(n-2)} \tag{2.33}$$

Similarly, we can obtain

$$\hat{p}_{dd} = \sum_{i=1}^{n} \frac{d_i(d_i - 1)}{n(n-1)(n-2)} \tag{2.34}$$

$$\hat{p}_{cd} = \sum_{i=1}^{n} \frac{a_i b_i' + a_i a_i' + b_i a_i' + b_i b_i'}{n(n-1)(n-2)} = \sum_{i=1}^{n} \frac{c_i d_i}{n(n-1)(n-2)} \tag{2.35}$$

Substituting the results (2.31) and (2.33) in (2.15), the estimated variance of T in samples for continuous variables is

$$n(n-1)\hat{\sigma}^2(T) = 8\hat{p}_c - 8\hat{p}_c^2(2n-3) + 16(n-2)\hat{p}_{cc}$$

$$n^2(n-1)^2\hat{\sigma}^2(T) = 8\left[2\sum_{i=1}^{n} c_i^2 - \frac{2n-3}{n(n-1)}\left(\sum_{i=1}^{n} c_i\right)^2 - \sum_{i=1}^{n} c_i\right] \tag{2.36}$$

In order to obviate any confusion regarding the calculation of the c_i and c_{ii} to estimate the variance from (2.36) in the case of no tied observations, a simple example is provided below for achievement tests in Mathematics and English administered to a group of six randomly chosen students.

Student	*A*	*B*	*C*	*D*	*E*	*F*
Math score	91	52	69	99	72	78
English score	89	72	69	96	66	67

The two sets of scores ranked and rearranged in order of increasing Mathematics scores are:

Student	*B*	*C*	*E*	*F*	*A*	*D*
Math rank	1	2	3	4	5	6
English rank	4	3	1	2	5	6

The numbers $c_i = a_i + b_i$ are

$$c_1 = 0+2 \quad c_2 = 0+2 \quad c_3 = 0+3 \quad c_4 = 1+2 \quad c_5 = 4+1 \quad c_6 = 5+0$$

$$\sum c_i = 20 \qquad \sum c_i^2 = 76 \qquad n = 6$$

$$\hat{p}_c = \frac{20}{6(5)} = \frac{2}{3}$$

$$\hat{p}_{cc} = \frac{76-20}{6(5)(4)} = \frac{7}{15}$$

$$t = 2(2/3) - 1 = 1/3$$

$$30^2\hat{\sigma}^2(T) = 8\left[2(76) - 20 - \frac{9}{6(5)}20^2\right] = 96$$

$$\hat{\sigma}^2(T) = 0.1067 \qquad \hat{\sigma}(T) = 0.33$$

If we wish to count the c_{ii} directly, we have for $c_{ii} = 2(\text{group } 1 + \text{group } 2 + \text{group } 3)$, the pairs relevant to c_{44}, say, are

Group 1: (1,5)(1,6)
Group 2: (5,6)
Group 3: None

so that $c_{44} = 2(3) = 6 = c_4(c_4 - 1)$.

On the other hand, suppose the English scores corresponding to increasing Math scores were ranked as

y	3	1	4	2	6	5

Then we can calculate

$$c_1 = c_4 = 3 \qquad c_2 = c_3 = c_5 = c_6 = 4$$

$$\hat{p}_c = 11/15 \qquad \hat{p}_{cc} = 1/2 \qquad t = 7/15 \qquad \hat{\sigma}^2(T) = -32/1125$$

and the estimated variance is negative! A negative variance from (2.15) of course cannot occur, but when the parameters p are replaced by estimates $\hat{p}$ and combined, the result can be negative. Since the probability estimates are consistent, the estimated variance of T will be positive for n sufficiently large.

Two applications of this asymptotic approximation to the nonnull distribution of T in nonparametric inference for large samples are:

1. An approximate $(1 - \alpha)100$ percent confidence-interval estimate of the population Kendall tau coefficient is

   $$t - z_{\alpha/2}\hat{\sigma}(T) < \tau < t + z_{\alpha/2}\hat{\sigma}(T)$$

2. An approximate test of

   $$H_0{:}\, \tau = \tau_0 \qquad \text{versus} \qquad H_1{:}\, \tau \neq \tau_0$$

   with significance level α is to reject H_0 when

   $$\frac{|t - \tau_0|}{\hat{\sigma}(T)} \geqslant z_{\alpha/2}$$

   A one-sided alternative can also be tested.

TIED OBSERVATIONS

Whether or not the marginal distributions of X and Y are assumed continuous, tied observations can occur within either or both samples. Ties across samples do not present any problem of course. Since the definition of A_{ij} in (2.3) assigned a value of zero to a_{ij} if a tie occurs in the (i,j) set of pairs for either the x or y sample values, T as defined before allows for, and essentially ignores, all zero differences. With τ defined as the difference $p_c - p_d, T$ as calculated from (2.6), (2.19), or (2.21) is an unbiased estimator of τ with variance as given in (2.13) even in the presence of ties. If the occurrence of ties in the sample is

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

attributed to a lack of precision in measurement as opposed to discrete marginal distributions, the simplified expression for var(T) in (2.15) may still be used. If there are sample ties, however, the expressions (2.20) and (2.22) are no longer equivalent to (2.6), (2.19), or (2.21).

For small samples with a small number of tied observations, the exact null distribution of T (or S) conditional on the observed ties can be determined by enumeration. There will be mw pairings of the two samples, each occurring with equal probability $1/mw$, if there are m and w distinguishable permutations of the x and y sample observations, respectively. For larger samples, the normal approximation to the distribution of T can still be used but with corrected moments. Conditional upon the observed ties, the parameters p_c, P_d, p_{cc}, p_{dd}, and p_{cd} must have a slightly different interpretation. For example, p_c and p_d here would be the probability that we select two pairs (x_i, y_i) and (x_j, y_j) which do not have a tie in either coordinate, and under the assumption of independence this is

$$\left[1-\frac{\sum u(u-1)}{n(n-1)}\right]\left[1-\frac{\sum v(v-1)}{n(n-1)}\right]$$

where u denotes the multiplicity of a tie in the x set and the sum is extended over all ties and v has the same interpretation for the y set. These parameters in the conditional distribution can be determined and substituted in (2.13) to find the conditional variance (see, for example, Noether, 1967, pp. 76–77). The conditional mean of T, however, is unchanged, since even for the new parameters we have $p_c = p_d$ for independent samples.

Conditional on the observed ties, however, there are not longer $\binom{n}{2}$ distinguishable sets of pairs to check for concordance, and thus if T is calculated in the ordinary way, it cannot equal one even for perfect agreement. Therefore an alternative definition of T in the presence of ties is to replace the $n(n-1)$ in the denominator of (2.6), (2.19), or (2.21) by a smaller quantity. To obtain a result still analogous to a correlation coefficient, we might take (2.20) as the definition of T in general. Since $\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}^2$ is the number of nonzero differences $X_j - X_i$ for all (i,j), the sum is the total number of distinguishable differences less the number involving tied observations, or $n(n-1) - \sum u(u-1)$. Similarly for the Y observations. Therefore our modified T from (2.20) is

$$T = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}V_{ij}}{\{[n(n-1) - \sum u(u-1)][n(n-1) - \sum v(v-1)]\}^{1/2}} \tag{2.37}$$

which reduces to all previously given forms if there are no ties. The modified T from (2.21) is

$$T = \frac{C - Q}{\left\{\left[\binom{n}{2} - \sum\binom{u}{2}\right]\left[\binom{n}{2} - \sum\binom{v}{2}\right]\right\}^{1/2}} \tag{2.38}$$

Note that the denominator in (2.38) is a function of the geometric mean of the number of untied X observations and the number of untied Y observations. The modified T in (2.37) or (2.38) is frequently called tau_b in order to distinguish it from (2.20) or (2.21), which is called tau_a and has no correction for ties.

The absolute value of the coefficient T calculated from (2.37) or (2.38) is always greater than the absolute of a coefficient calculated from (2.20) or (2.21) when ties are present, but it still may not be equal to one for perfect agreement or disagreement. The only way to define a tau coefficient that does always equal one for perfect agreement or disagreement is to define

$$\gamma = \frac{C - Q}{C + Q} \tag{2.39}$$

This ratio, the number of concordant pairs with no ties minus the number of discordant pairs with no ties by the total number of untied pairs, is called the *Goodman-Kruskal gamma coefficient*.

A RELATED MEASURE OF ASSOCIATION FOR DISCRETE POPULATIONS

In Section 11.1 we stated the criterion that a good measure of association between two random variables would equal $+1$ for a perfect direct relationship and -1 for a perfect indirect relationship. In terms of the probability parameters, perfect concordance requires $p_c = 1$, and perfect discordance requires $p_d = 1$. With Kendall's coefficient defined as $\tau = p_c - p_d$, the criterion is satisfied if and only if $p_c + p_d = 1$. But if the marginal distributions of X and Y are not continuous,

$$\begin{aligned} p_c + p_d &= P[(X_j - X_i)(Y_j - Y_i) > 0] + P[(X_j - X_i)(Y_j - Y_i) < 0] \\ &= 1 - P[(X_j - X_i)(Y_j - Y_i) = 0] \\ &= 1 - P[(X_i = X_j) \cup (Y_i = Y_j)] = 1 - p_t \end{aligned}$$

where p_t denotes the probability that a pair is neither concordant not discordant. Thus τ cannot be considered a "good" measure of association if $p_t \neq 0$.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

However, a modified parameter which does satisfy the criteria for all distributions can easily be defined as

$$\tau^* = \frac{\tau}{1 - p_t} = p_c^* - p_d^*$$

where p_c^* and p_d^* are, respectively, the conditional probabilities of concordance and discordance given that there are no ties

$$p_c^* = \frac{p_c}{1 - p_t} = \frac{P[(X_j - X_i)(Y_j - Y_i) > 0]}{P[(X_j - X_i)(Y_j - Y_i) \neq 0]}$$

Since τ^* is a linear function of τ, an estimate is provided by

$$T^* = \frac{T}{1 - \hat{p}_t} = \frac{\hat{p}_c - \hat{p}_d}{\hat{p}_c + \hat{p}_d}$$

with $\hat{p}_c$ and $\hat{p}_d$ defined as before in (2.31) and (2.32). Since $\hat{p}_c$ and $\hat{p}_d$ are consistent estimators, the asymptotic distribution of $T/(\hat{p}_c + \hat{p}_d)$ is equivalent to the asymptotic distribution of $T/(p_c + p_d)$, which we know to be the normal distribution. Therefore for large samples, inferences concerning τ^* can be made (see, for example, Goodman and Kruskal, 1954, 1959, 1963).

USE OF KENDALL'S STATISTIC TO TEST AGAINST TREND

In Chapter 3 regarding tests for randomness, we observed that the arrangement of relative magnitudes in a single sequence of time-ordered observations can indicate some sort of trend. When the theory of runs up and down was used to test a hypothesis of randomness, the magnitude of each observations relative to its immediately preceding value was considered, and a long run of plus (minus) signs or a sequence with a large predominance of plus (minus) signs was considered indicative of an upward (downward) trend. If time is treated as an X variable, say, and a set of time-ordered observations as the Y variable, an association between X and Y might be considered indicative of a trend. Thus the degree of concordance between such X and Y observations would be a measure of trend, and Kendall's tau statistic becomes a measure of trend. Unlike the case of runs up and down, however, the tau coefficient considers the relative magnitude of each observation relative to *every* preceding observation.

A hypothesis of randomness in a single set of n time-ordered observations is the same as a hypothesis of independence between these observations when paired with the numbers $1, 2, \ldots, n$. Therefore,

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

assuming that $x_i = i$ for $i = 1, 2, \ldots, n$, the indicator variables A_{ij} defined in (2.3) become

$$A_{ij} = \operatorname{sgn}(j - 1)\operatorname{sgn}(Y_j - Y_i)$$

and (2.6) can be written as

$$\binom{n}{2} T = \sum\sum_{1 \leqslant i < j \leqslant n} \operatorname{sgn}(Y_j - Y_i)$$

The exact null distribution of T is the same as before. If the alternative is an upward trend, the rejection region consists of large positive values of T, and T can be considered an unbiased estimate of τ, a relative measure of population trend. For a downward trend, we reject for large negative values of T. This use of T is frequently called the *Mann Test*.

11.3 SPEARMAN'S COEFFICIENT OF RANK CORRELATION

A random sample of n pairs

$$(X_1, Y_1),\ (X_2, Y_2), \ldots, (X_n, Y_n)$$

is drawn from a bivariate population with Pearson product-moment correlation coefficient ρ. In classical statistics, the estimate commonly used for ρ is the *sample correlation coefficient* defined as

$$R = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2)\right]^{1/2}} \tag{3.1}$$

In general, of course, the sampling distribution of R depends upon the form of the bivariate population from which the sample of pairs is drawn. However, suppose the X observations are ranked from smallest to largest using the integers $1, 2, \ldots, n$, and the Y observations are ranked separately using the same ranking scheme. In other words, each observation is assigned a rank according to its magnitude relative to the others in its own group. If the marginal distributions of X and Y are assumed continuous, unique sets of rankings exist theoretically. The data then consist of n sets of paired ranks from which R as defined in (3.1) can be calculated. The resulting statistic is then called *Spearman's coefficient of rank correlation*. It measures the degree of correspondence between rankings, instead of between actual variate values, but it can still be considered a measure of association between the samples and an estimate of the association between X and Y in the continuous bivariate population. It is difficult to interpret exactly what R is estimating in the population from which these samples were drawn and ranks obtained, but the measure has intuitive appeal anyway. The problem of interpretation will be treated in Section 11.4.

The fact that we know the numerical values of the derived observations from which Spearman's R is computed, if not their scheme of pairing, means that the expression in (3.1) can be considerably simplified. Denoting the respective ranks of the random variables in the samples by

$$R_i = \text{rank}(X_i) \qquad \text{and} \qquad S_i = \text{rank}(Y_i)$$

the derived sample observations of n pairs are

$$(r_1, s_1), (r_2, s_2), \ldots, (r_n, s_n)$$

Since addition is commutative, we have the constant values for all samples

$$\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \qquad \bar{R} = \bar{S} = \frac{n+1}{2} \tag{3.2}$$

$$\sum_{i=1}^{n} (R_i - \bar{R})^2 = \sum_{i=1}^{n} (S_i - \bar{S})^2 = \sum_{i=1}^{n} \left(i - \frac{(n+1)}{2} \right)^2 = \frac{n(n^2-1)}{12} \tag{3.3}$$

Substituting these constants in (3.1), the following equivalent forms of R are obtained:

$$R = \frac{12 \sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})}{n(n^2-1)} \tag{3.4}$$

$$R = \frac{12 \left[\sum_{i=1}^{n} R_i S_i - n(n+1)^2/4 \right]}{n(n^2-1)} \tag{3.5}$$

$$R = \frac{12 \sum_{i=1}^{n} R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1} \tag{3.6}$$

Another useful form of R is in terms of the differences.

$$D_i = R_i - S_i = (R_i - \bar{R}) - (S_i - \bar{S})$$

Substituting (3.3) in the expression

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (R_i - \bar{R})^2 + \sum_{i=1}^{n} (S_i - \bar{S})^2 - 2 \sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})$$

and using the result back in (3.4), the most common form of the Spearman coefficient of rank correlation is obtained as

$$R = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \tag{3.7}$$

We can assume without loss of generality that the n sample pairs are labeled in accordance with increasing magnitude of the X component, so that $R_i = i$ for $i = 1, 2, \ldots, n$. Then S_i is the rank of the Y observation that is paired with the rank i in the X sample, and $D_i = i - S_i$.

In Section 11.1, criteria were defined for a "good" relative measure of association between two random variables. Although the parameter analogous to R has not been specifically defined, we can easily verify that Spearman's R does satisfy the corresponding criteria of a good measure of association between sample ranks.

1. For any two sets of paired ranks (i, S_i) and (j, S_j) of random variables in a sample from any continuous bivariate distribution, in order to have perfect concordance between ranks, the Y component must also be increasing, or, equivalently, $s_i = i$ and $d_i = 0$ for $i = 1, 2, \ldots, n$ so that R equals 1.
2. For perfect discordance between ranks, the Y arrangement must be the reverse of the X arrangement to have decreasing Y components, so that $s_i = n - i + 1$ and

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [i - (n - i + 1)]^2 = 4\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{3}$$

from (3.3). Substituting this in (3.7), we find $R = -1$.

3–6. Since R in, say, (3.7) is algebraically equivalent to (3.1) and the value of (3.1) is in the interval $[-1, 1]$ for all sets of numerical pairs, the same bounds apply here. Further, R is commutative and symmetric about zero and has expectation zero when the X and Y observations are independent. These properties will be shown later in the section.

7. Since ranks are preserved under all order-preserving transformations, the measure R based on ranks is invariant.

EXACT NULL DISTRIBUTION OF *R*

If the X and Y random variables from which these n pairs of ranks (R_i, S_i) are derived are independent, R is a distribution-free statistic since each of the $n!$ distinguishable sets of pairing of n ranks is equally likely. Therefore, the random sampling distribution of R can be determined and the statistic used to perform exact distribution-free tests of independence. If we let u_r denote the number of pairings which lead to a value r for the statistic, the null probability distribution is

$$f_R(r) = \frac{u_r}{n!}$$

The null distribution of R is symmetric about the origin, since the random variable $D = \sum_{i=1}^{n} D_i^2$ is symmetric about $n(n^2-1)/6$. This property is the result of the fact that for any sets of pairs

$$(1, s_1),\ (2, s_2), \ldots, (n, s_n)$$

with

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (i - s_i)^2$$

there exists a conjugate set of pairs

$$(1, s_n),\ (2, s_{n-1}), \ldots, (n, s_1)$$

with

$$\sum_{i=1}^{n} d_i'^2 = \sum_{i=1}^{n} (i - s_{n-i+1})^2 = \sum_{i=1}^{n} (n - i + 1 - s_i)^2$$

The sums of squares of the respective sum and difference of rank differences are

$$\sum_{i=1}^{n} (d_i + d_i')^2 = \sum_{i=1}^{n} (n + 1 - 2s_i)^2 = 4\sum_{i=1}^{n} \left(s_i - \frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{3}$$

$$\sum_{i=1}^{n} (d_i - d_i')^2 = \sum_{i=1}^{n} (2i - n - 1)^2 = 4\sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{3}$$

Substituting these results in the relation

$$\sum_{i=1}^{n} [(d_i + d_i') + (d_i - d_i')]^2$$

$$= 4\sum_{i=1}^{n} {d_i}^2 = \sum_{i=1}^{n} (d_i + d_i')^2 + \sum_{i=1}^{n} (d_i - d_i')^2 + 2\sum_{i=1}^{n} ({d_i}^2 - d_i'^2)$$

we obtain

$$4\sum_{i=1}^{n} d_i^2 = \frac{2n(n^2-1)}{3} + 2\sum_{i=1}^{n} d_i^2 - 2\sum_{i=1}^{n} d_i'^2$$

or

$$\sum_{i=1}^{n} d_i^2 + \sum_{i=1}^{n} d_i'^2 = \frac{n(n^2-1)}{3} = \text{constant}$$

Further, R cannot equal zero unless n is even, since $\sum_{i=1}^{n} d_i^2$ is always even because $\sum_{i=1}^{n} d_i = 0$, an even number.

The direct approach to determining u_r is by enumeration, which is probably least tedious for R in the form of (3.6). Because of the symmetry property, only $n!/2$ cases need be considered. For $n = 3$, for example, we list the following sets (s_1, s_2, s_3) which may be paired with (1,2,3), and the resulting values of R.

(s_1, s_2, s_3)	$\sum_{i=1}^{n} is_i$	r
$1, 2, 3$	14	1.0
$1, 3, 2$	13	0.5
$2, 1, 3$	13	0.5

The complete probability distribution then is

r	-1.0	-0.5	0.5	1.0
$f_R(r)$	1/6	2/6	2/6	1/6

This method of generating the distribution is time consuming, even for moderate n. Of course, there are more efficient methods of enumeration (see, for example, Kendall and Gibbons, 1990, pp. 97–98). The probability distribution of R is given in Table M of the Appendix as tail probabilities for $n \leqslant 10$ and as critical values for $11 \leqslant n \leqslant 30$. More extensive tables of the exact null distribution of R or $\sum D^2$ are given in Glasser and Winter (1961), Owen (1962), De Jonge and Van Montfort (1972), Zar (1972), Otten (1973a,b) Dunstan, Nix, and Reynolds (1979), Neave (1981), Nelson (1986), Franklin (1988b), Ramsay (1989), and Kendall and Gibbons (1990).

Although the general null probability distribution of R requires enumeration, the marginal and joint distributions of any number of the individual ranks of a single random sample of size n are easily determined from combinatorial theory. For example, for the Y sample, we have

$$f_{S_i}(s_i) = \frac{1}{n} \qquad s_i = 1, 2, \ldots, n \tag{3.8}$$

$$f_{S_i,S_j}(s_i, s_j) = \frac{1}{n(n-1)} \qquad s_i, s_j = 1, 2, \ldots, n, s_i \neq s_j \tag{3.9}$$

Thus, using (3.2) and (3.3),

$$E(S_i) = \frac{n+1}{2} \qquad \operatorname{var}(S_i) = \frac{n^2-1}{12}$$

For the covariance, we have for all $i \neq j$,

$$\begin{aligned}
\operatorname{cov}(S_i, S_j) &= E(S_iS_j) - E(S_i)E(S_j) \\
&= \frac{1}{n(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} ij - \frac{1}{n^2}\left(\sum_{i=1}^{n} i\right)^2 \\
&= \frac{1}{n^2(n-1)}\left[n\left(\sum_{i=1}^{n} i\right)^2 - n\sum_{i=1}^{n} i^2 - (n-1)\left(\sum_{i=1}^{n} i\right)^2\right] \\
&= \frac{-1}{n^2(n-1)}\left[\frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}\right] = -\frac{n+1}{12}
\end{aligned} \tag{3.10}$$

The same results hold for the ranks R_i of the X sample. Under the null hypothesis that the X and Y samples are independent, the ranks R_i and S_j are independent for all i, j, and the null mean and variance of R easily found as follows:

$$E\left(\sum_{i=1}^{n} R_iS_i\right) = nE(R_i)E(S_i) = \frac{n(n+1)^2}{4} \tag{3.11}$$

$$\begin{aligned}
\operatorname{var}\left(\sum_{i=1}^{n} R_iS_i\right) &= n \operatorname{var}(R_i) \operatorname{var}(S_i) + n(n-1) \operatorname{cov}(R_i, R_j) \operatorname{cov}(S_i, S_j) \\
&= \frac{n(n^2-1)^2 + n(n-1)(n+1)^2}{144} = \frac{n^2(n-1)(n+1)^2}{144}
\end{aligned} \tag{3.12}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Then using the form of R in (3.6)

$$E(R \mid H_0) = 0 \qquad \text{var}(R \mid H_0) = \frac{1}{n-1} \tag{3.13}$$

ASYMPTOTIC NULL DISTRIBUTION OF *R*

Considering R in the form of (3.6), and as before assuming S_i denotes the rank of the Y observation paired with the ith smallest X observation, we see that the distribution of R depends only on the random variables $\sum_{i=1}^{n} iS_i$. This quantity is a linear combination of random variables, which can be shown to be asymptotically normally distributed (see, for example, Fraser, 1957, pp. 247–248). The mean and variance are given in (3.11) and (3.12). The standardized normal variable used for an approximate test of independence then is

$$Z = \left(12\sum_{i=1}^{n} iS_i - 3n^3\right) n^{-5/2}$$

or, equivalently,

$$Z = R\sqrt{n-1} \tag{3.14}$$

There is some disagreement in the literature about the accuracy of this approximation for moderate n. Some authors claim that the statistic

$$t = \frac{R\sqrt{n-2}}{1-R^2} \tag{3.15}$$

which has approximately Student's t distribution with n$-$2 degrees of freedom, gives more accurate results for moderate n.

TESTING THE NULL HYPOTHESIS

Since R has mean zero for independent random variables, the appropriate rejection region of size α is large absolute values of R for a general alternative of nonindependence and large positive values of R for alternatives of positive dependence. As in the case of Kendall's coefficient, if the null hypothesis of independence is accepted, we can infer that $\rho(X, Y)$ equals zero, but dependence between the variables does not necessarily imply that $\rho(X, Y) \neq 0$. Besides, the coefficient of rank correlation is measuring association between ranks, not variate values. Since the distribution of R was derived only under the assumption of independence, these results cannot be used to construct confidence-interval estimates of $\rho(X, Y)$ or $E(R)$.

TIED OBSERVATIONS

In all of the foregoing discussion we assumed that the data to be analyzed consisted of n sets of paired integer ranks. These integer ranks may be obtained by ordering observations from two continuous population, but the theory is also equally applicable to any two sets of n pairs which can be placed separately in a unique preferential order. In the first case, ties can still occur within either or both sets of sample measurements, and in the second case it is possible that no preference can be made between two or more of the individuals in either group. Thus, for practical purposes, the problem of ties within a set of ranks must be considered.

If within each set of tied observations the ranks they would have if distinguishable are assigned at random, nothing is changed since we still have the requisite type of data to be analyzed. However, such an approach has little intuitive appeal, and besides an additional element of chance is introduced. The most common practice for dealing with tied observations here, as in most other nonparametric procedures, is to assign equal ranks to indistinguishable observations. If that rank is the midrank in every case, the sum of the ranks for each sample is still $n(n+1)/2$, but the sum of squares of the ranks is changed, and the expressions in (3.4) to (3.7) are no longer equivalent to (3.1). Assuming that the spirit of the rank correlation coefficient is unchanged, the expressions in (3.1) can be calculated directly from the ranks assigned. However, a form analogous to (3.7) which is equivalent to (3.1) can still be found for use in the presence of ties.

We shall investigate what happens to the sum of squares

$$\sum_{i=1}^{n}(s_i-\bar{s})^2=\sum_{i=1}^{n}s_i^2-\frac{n(n+1)^2}{2}$$

when there are one or more groups of u tied observations within the Y sample and each is assigned the appropriate midrank. In each group of u tied observations which, if not tied, would be assigned the ranks $p_k+1, p_k+2, \ldots, p_k+u$, the rank assigned to all is

$$\sum_{i=1}^{n}\frac{p_k+i}{u}=p_k+\frac{u+1}{2}$$

The sum of squares for these tied ranks then is

$$u\left(p_k+\frac{u+1}{2}\right)^2=u\left[p_k^2+p_k(u+1)+\frac{(u+1)^2}{4}\right] \tag{3.16}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and the corresponding sum in the absence of ties would be

$$\sum_{i=1}^{u}(p_k+i)^2 = up_k^2 + p_k u(u+1) + \frac{u(u+1)(2u+1)}{6} \tag{3.17}$$

This particular group of u tied observations then decreases the sum of squares by the difference between (3.17) and (3.16) or

$$\frac{u(u+1)(2u+1)}{6} - \frac{u(u+1)^2}{4} = \frac{u(u^2-1)}{12}$$

Since this is true for each group of u tied observations, the sum of squares in the presence of ties is

$$\sum_{i=1}^{n}(s_i-\bar{s})^2 = \frac{n(n^2-1)}{12} - u' \tag{3.18}$$

where $u' = \sum u(u^2-1)/12$ and the summation is extended over all sets of u tied ranks in the Y sample. Letting t' denote the corresponding sum for the X sample, we obtain the alternative forms of (3.1) as

$$R = \frac{12[\sum_{i=1}^{n} R_iS_i - n(n+1)^2/4]}{\{[n(n^2-1)-12t'][n(n^2-1)-12u']\}^{1/2}} \tag{3.19}$$

or

$$R = \frac{n(n^2-1) - 6\sum_{i=1}^{n} D_i^2 - 6(t'+u')}{\{[n(n^2-1)-12t'][n(n^2-1)-12u']\}^{1/2}} \tag{3.20}$$

analogous to (3.5) and (3.7), respectively, since here

$$\sum_{i=1}^{n} D_i^2 = \frac{n(n^2-1)}{6} - t' - u' - 2\sum_{i=1}^{n}(R_i-\bar{R})(S_i-\bar{S})$$

Assuming this to be our definition of the sample coefficient of rank correlation in the presence of ties, its probability distribution under the null hypothesis of independence is clearly not the same as the null distribution discussed before for n distinct ranks. For small n, it is possible again to obtain by enumeration the exact null distribution conditional upon a given set of ties. This of course is very tedious. The asymptotic distribution of our R as modified for ties is also normal since it is still a linear combination of the S_i random variables. Since the total sum of ranks is unchanged when tied ranks are assigned by the midrank method, $E(S_i)$ is unchanged and $E(R \mid H_0)$ is obviously still zero. The fact that the variance of modified R is also unchanged in the presence of ties is not so obvious. The marginal and joint distributions of the ranks of the Y sample in the presence of ties can still be written in the forms (3.8) and (3.9) except that that domain is now n numbers, not all distinct, which we can write as $s'_1, s'_2, \ldots, s'_n$. Then using (3.18),

$$\mathrm{var}(S_i) = \sum_{i=1}^{n} \frac{(s'_i - \bar{s})^2}{n} = \frac{n(n^2-1) - 12u'}{12n}$$

For the covariance, proceeding as in the steps leading to (3.10),

$$\mathrm{cov}(S_i, S_j) = \frac{1}{n(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} s'_i s'_j - \bar{s}^2$$

$$= -\frac{\sum_{i=1}^{n} (s'_i - \bar{s})^2}{n(n-1)} = -\frac{n(n^2-1) - 12u'}{12n(n-1)}$$

Similar results hold for the X ranks. Now using R in the form of (3.19), we have

$$\mathrm{var}(R \mid H_0) = \frac{144[n\mathrm{var}(R_i)\,\mathrm{var}(S_i) + n(n-1)\,\mathrm{cov}(R_i, R_j)\,\mathrm{cov}(S_i, S_j)]}{[n(n^2-1) - 12t'][n(n^2-1) - 12u']}$$

and substitution of the appropriate variances and covariances gives as before

$$\mathrm{var}(R \mid H_0) = \frac{1}{n-1}$$

Thus for large samples with ties, a modified $R\sqrt{n-1}$ with R calculated from (3.19) or (3.20) can still be treated as a standard normal variable for testing a hypothesis of independence. However, unless the ties are extremely extensive, they will have little effect on the value of R. In practice, the common expression given in (3.7) is often used without corrections for ties. It should be noted that the effect of the correction factors is to decrease the value of R. This means that a negative R is closer to -1, not to zero.

USE OF SPEARMAN'S *R* TO TEST AGAINST TREND

As with Kendall's T, R can be considered a measure of trend in a single sequence of time-ordered observations and used to test a null hypothesis of no trend. This application is called *Daniels' test*.

11.4 THE RELATIONS BETWEEN *R* AND *T*; *E*(*R*), τ, AND ρ

In Section 11.1 we defined the parameters τ and ρ as two different measures of association in a bivariate population, one in terms of concordances and the other as a function of covariance, but noted that concordance and covariance measure relationship in the same spirit at least. The sample estimate of τ was found to have exactly the same numerical value and theoretical properties whether calculated in terms of actual variate values or ranks, since the parameter τ and its estimate are both invariant under all order-preserving transformations. However, this is not true for the parameter ρ or for a sample estimate calculated from (3.1) with variate values. The Pearson product-moment correlation coefficient is invariant under linear transformations only, and ranks usually cannot be generated using only linear transformations.

The coefficient of rank correlation is certainly a measure of association between ranks. It has a certain intuitive appeal as an estimate of ρ, but it is not a direct sample analog of this parameter. Nor can it be considered a direct sample analog of a "population coefficient of rank correlation" if the marginal distributions of our random variables are continuous, since theoretically continuous random variables cannot be ranked. If an infinite number of values can be assumed by a random variable, the values cannot be enumerated and therefore cannot be ordered. However, we still would like some conception, however nebulous, of a population parameter which is the analog of the Spearman coefficient of rank correlation in a random sample of pairs from a continuous bivariate population. Since probabilities of order properties are population parameters and these probabilities are the same for either ranks or variate values, if R can be defined in terms of sample proportions of types of concordance, as T was, we shall be able to define a population parameter other than ρ for which the coefficient of rank correlation is an unbiased estimate.

For this purpose, we first investigate the relationship between R and T for samples with no ties from any continuous bivariate population. In (2.20), T was written in a form resembling R as

$$T = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{U_{ij}V_{ij}}{n(n-1)} \tag{4.1}$$

where

$$U_{ij} = \operatorname{sgn}(X_j - X_i) \qquad \text{and} \qquad V_{ij} = \operatorname{sgn}(Y_j - Y_i) \tag{4.2}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

To complete the similarity, we must determine the general relation between U_{ij} and R_i, V_{ij}, and S_i. A functional definition of R_i was given in (5.5.1) as

$$R_i = 1 + \sum_{\substack{j=1 \\ j \neq i}}^{n} S(X_i - X_j) \tag{4.3}$$

where

$$S(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u \geqslant 0 \end{cases}$$

In general, then the relation is

$$\begin{aligned} &\operatorname{sgn}(X_j - X_i) = 1 - 2S(X_i - X_j) \qquad \text{for all } 1 \leqslant i \neq j \leqslant n \\ &\operatorname{sgn}(X_i - X_i) = 0 \end{aligned} \tag{4.4}$$

Substituting this form back in (4.3), we have

$$R_i = \frac{n+1}{2} - \frac{1}{2}\sum_{j=1}^{n} \operatorname{sgn}(X_j - X_i)$$

or

$$R_i - \bar{R} = -\sum_{j=1}^{n} \frac{U_{ij}}{2}$$

Using R in the form (3.4), by substitution we have

$$\begin{aligned} n(n^2-1)R &= 12\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S}) = 3\sum_{i=1}^{n}\left(\sum_{j=1}^{n} U_{ij} \sum_{k=1}^{n} V_{ik}\right) \\ &= 3\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}V_{ij} + 3\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{k=1 \\ k \neq j}}^{n} U_{ij}V_{ik} \end{aligned}$$

or from (4.1)

$$R = \frac{3}{n+1}T + \frac{6}{n(n^2-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i<j}}^{n}\sum_{\substack{k=1 \\ k \neq j}}^{n} U_{ij}V_{ik} \tag{4.5}$$

Before, we defined two pairs (X_i, Y_i) and (X_j, Y_j) as being concordant if $U_{ij}V_{ij} > 0$, with p_c denoting the probability of concordance

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and $\hat{p}_c$ the corresponding sample estimate, the number of concordant sample pairs divided by $n(n-1)$, the number of distinguishable pairs, and found $T = 2\hat{p}_c - 1$. To complete a definition of R in terms of concordances, because of the last term in (4.5), we must now define another type of concordance, this time involving three pairs. We shall say that the three pairs (X_i, Y_i), (X_j, Y_j) and (X_k, Y_k) exhibit a concordance of the second order if

$$X_i < X_j \text{ whenever } Y_i < Y_k$$

or

$$X_i > X_j \text{ whenever } Y_i > Y_k$$

or, equivalently, if

$$(X_j - X_i)(Y_k - Y_i) = U_{ij}V_{ik} > 0$$

The probability of a second-order concordance is

$$p_{c_2} = P[(X_j - X_i)(Y_k - Y_i) > 0]$$

and the corresponding sample estimate $\hat{p}_{c_2}$ is the number of sets of three pairs with the product $U_{ij}V_{ik} > 0$ for $i < j$, $k \neq j$, divided by $\binom{n}{2}(n-2)$, the number of distinguishable sets of three pairs. The triple sum in (4.5) is the totality of all these products, whether positive or negative, and therefore equals

$$\binom{n}{2}(n-2)[\hat{p}_{c_2} - (1 - \hat{p}_{c_2})] = \frac{n(n-1)(n-2)(2\hat{p}_{c_2} - 1)}{2}$$

In terms of sample concordances, then, (4.5) can be written as

$$R = \frac{3}{n+1}(2\hat{p}_c - 1) + \frac{3(n-2)}{n+1}(2\hat{p}_{c_2} - 1) \tag{4.6}$$

and the population parameter for which R is an unbiased estimator is

$$E(R) = \frac{3[\tau + (n-2)(2p_{c_2} - 1)]}{n+1} \tag{4.7}$$

We shall now express p_{c_2} for any continuous bivariate population $F_{X,Y}(x, y)$ in a form analogous to (2.17) for p_c:

$$
\begin{aligned}
p_{c_2} &= P[(X_i < X_j) \cap (Y_i < Y_k)] + P[(X_i > X_j) \cap (Y_i > Y_k)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{P[(X_i < x_j) \cap (Y_i < y_k)] \\
&\quad + P[(X_i > x_j) \cap (Y_i > y_k)]\} f_{X_j,Y_k}(x_j, y_k)\, dx_j\, dy_k \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [F_{X,Y}(x,y) + 1 - F_X(x) - F_Y(y) + F_{X,Y}(x,y)]\, dF_X(x)\, dF_Y(y) \\
&= 1 + 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{X,Y}(x,y)\, dF_X(x)\, dF_Y(y) - 2\int_{-\infty}^{\infty} F_X(x)\, dF_X(x) \\
&= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{X,Y}(x,y)\, dF_X(x)\, dF_Y(y) \qquad (4.8)
\end{aligned}
$$

A similar development yields another equivalent form

$$p_{c_2} = 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_X(x) F_Y(y)\, dF_{X,Y}(x,y) \qquad (4.9)$$

If X and Y are independent, of course, a comparison of these expressions with (2.17) shows that $p_{c_2} = p_c = 1/2$. Unlike p_c, however, which ranges between 0 and 1, p_{c_2} ranges only between 1/3 and 2/3, with the extreme values obtained for perfect indirect and direct linear relationships, respectively. This result can be shown easily. For the upper limit, since for all x, y,

$$2F_X(x)F_Y(y) \leqslant F_X^2(x) + F_Y^2(y)$$

we have from (4.9)

$$p_{c_2} \leqslant 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_X^2(x)\, dF_{X,Y}(x,y) = 2/3$$

Similarly, for all x, y,

$$2F_X(x)F_Y(y) = [F_X(x) + F_Y(y)]^2 - F_X^2(x) - F_Y^2(y)$$

so that from (4.9)

$$
\begin{aligned}
p_{c_2} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [F_X(x) + F_Y(y)]^2\, dF_{X,Y}(x,y) - 2/3 \\
&\geqslant \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [F_X(x) + F_Y(y)]\, dF_{X,Y}(x,y)\right]^2 - 2/3 = 1/3
\end{aligned}
$$

Now if X and Y have a perfect direct linear relationship, we can assume without loss of generality that $X = Y$, so that

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$F_{X,Y}(x,y) = \begin{cases} F_X(x) & \text{if } x \leqslant y \\ F_X(y) & \text{if } x > y \end{cases}$$

Then from (4.8)

$$p_{c_2} = 2(2)\int_{-\infty}^{\infty}\int_{-\infty}^{y} F_X(x)f_X(x)f_X(y)\,dx\,dy = 2/3$$

For a perfect indirect relationship, we assume $X = -Y$, so that

$$F_{X,Y}(x,y) = \begin{cases} F_X(x) - F_X(-y) & \text{if } x \geqslant -y \\ 0 & \text{if } x < -y \end{cases}$$

and

$$\begin{aligned} p_{c_2} &= 2\int_{-\infty}^{\infty}\int_{-y}^{\infty} [F_X(x) - F_X(-y)]f_X(x)f_X(-y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} \{1 - F_X^2(-y) - 2[1 - F_X(-y)]F_X(-y)\}f_X(-y)\,dy \\ &= \int_{-\infty}^{\infty} [1 - F_X(-y)]^2 f_X(-y)\,dy = 1/3 \end{aligned}$$

Substitution of these extreme values in (4.7) shows that for any continuous population ρ, τ, and $E(R)$ all have the same value for the following cases:

X,Y relation	$\rho = \tau = E(R)$
Indirect linear dependence	-1
Independence	0
Direct linear dependence	1

Although strictly speaking we cannot talk about a parameter for a bivariate distribution which is a coefficient of rank correlation, it seems natural to define the pseudo rank-correlation parameter, say ρ_2, as that constant for which R is an unbiased estimator in large samples. Then from (4.7), we have the definition

$$\rho_2 = \lim_{n\to\infty} E(R) = 3(2p_{c_2} - 1) \tag{4.10}$$

and for a sample of size n, the relation between $E(R)$, ρ_2, and τ is

$$E(R) = \frac{3\tau + (n-2)\rho_2}{n+1} \tag{4.11}$$

The relation between ρ_2 (for ranks) and ρ (for variate values) depends on the relation between p_{c_2} and covariance. From (4.9), we see that

$$\begin{aligned} p_{c_2} &= 2E[F_X(X)F_Y(Y)] = 2\,\text{cov}[F_X(X), F_Y(Y)] + 2E[F_X(X)]E[F_Y(Y)] \\ &= 2\,\text{cov}[F_X(X), F_Y(Y)] + 1/2 \end{aligned}$$

Since

$$\text{var}[F_X(X)] = \text{var}[F_Y(Y)] = 1/12$$

we have

$$6p_{c_2} = \rho[F_X(X), F_Y(Y)] + 3$$

and we see from (4.10) that

$$\rho_2 = \rho[F_X(X), F_Y(Y)]$$

Therefore ρ_2 is sometimes called the *grade correlation coefficient*, since the grade of a number x is usually defined as the cumulative probability $F_X(x)$.

11.5 ANOTHER MEASURE OF ASSOCIATION

Another nonparametric type of measure of association for paired samples which is related to the Pearson product-moment correlation coefficient has been investigated by Fieller, Hartley, Pearson, and others. This is the ordinary Pearson sample correlation coefficient of (3.1) calculated using expected normal scores in place of ranks or variate values. That is, if $\xi_i = E(U_{(i)})$, where $U_{(i)}$ is the ith order statistic in a sample of n from the standard normal population and S_i denotes the rank of the Y observation which is paired with the ith smallest X observation, the random sample of pairs of ranks

$$(1, s_1), (2, s_2), \ldots, (n, s_n)$$

is replaced by the derived sample of pairs

$$(\xi_1, \xi_{s_1}), (\xi_2, \xi_{s_2}), \ldots, (\xi_n, \xi_{s_n})$$

and the correlation coefficient for these pairs is

$$R_F = \frac{\sum_{i=1}^{n} \xi_i \xi_{s_i}}{\sum_{i=1}^{n} \xi_i^2}$$

This coefficient is discussed in Fieller, Hartley, and Pearson (1957) and Fieller and Pearson (1961). The authors show that the transformed random variable

$$Z_F = \tanh^{-1} R_F$$

is approximately normally distributed with moments

$$E(Z_F) = \tanh^{-1}\left[\rho\left(1 - \frac{0.6}{n+8}\right)\right]$$

$$\text{var}(Z_F) = \frac{1}{n-3}$$

where ρ is the correlation coefficient in the bivariate population from which the sample is drawn.

The authors also show that analogous transformations on R and T,

$$Z_R = \tanh^{-1} R$$

$$Z_T = \tanh^{-1} T$$

produce approximately normally distributed random variables, but in the nonnull case the approximation for Z_F is best.

11.6 APPLICATIONS

Kendall's sample tau coefficient (Section 11.2) is one descriptive measure of association in a bivariate sample. The statistic is calculated as

$$T = \frac{2S}{n(n-1)} = \frac{2(C-Q)}{n(n-1)}$$

where C is the number of concordant pairs and Q is the number of discordant pairs among (X_i, Y_i) and (X_j, Y_j), for all $i < j$ in a sample of n observations. T ranges between -1 and 1, with -1 describing perfect disagreement, 1 describing perfect agreement, and 0 describing no agreement. The easiest way to calculate C and Q is to first arrange one set of observations in an array, while keeping the pairs intact. A pair in which there is a tie in either the X observations or the Y observations is not counted as part of either C or Q, and therefore with ties it may be necessary to list all possible pairs to find the correct values for C and Q. The modified T is then calculated from (2.37) and called tau$_b$.

The null hypothesis of independence between X and Y can be tested using T. The appropriate rejection regions and P values for an observed value t are as follows:

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Alternative	*Rejection region*	*P value*
Positive dependence	$T \geqslant t_\alpha$	$P(T \geqslant t)$
Negative dependence	$T \leqslant -t_\alpha$	$P(T \leqslant t)$
Nonindependence	$T \geqslant t_{\alpha/2}$ or $T \leqslant -t_{\alpha/2}$	2 (smaller of above)

The exact cumulative null distribution of T is given in Table L of the Appendix as right-tail probabilities for $n \leqslant 10$. Quantiles of T are also given for $11 \leqslant n \leqslant 30$. For $n > 30$, the normal approximation to the null distribution of T indicates the following rejection regions and P values:

Alternative	*Rejection region*	*P value*
Positive dependence	$T \geqslant z_\alpha \sqrt{2(2n+5)}/3\sqrt{n(n-1)}$	$P(z \geqslant 3t\sqrt{n(n-1)}/\sqrt{2(2n+5)})$
Negative dependence	$T \leqslant -z_\alpha \sqrt{2(2n+5)}/3\sqrt{n(n-1)}$	$P(z \leqslant 3t\sqrt{n(n-1)}/\sqrt{2(2n+5)})$
Nonindependence	Both above with $z_{\alpha/2}$	2 (smaller of above)

This test of the null hypothesis of independence can also be used for the alternative of a trend in a time-ordered sequence of observations Y if time is regarded as X. The alternative of an upward trend corresponds to the alternative of positive dependence. This use of Kendalls's tau is frequently called the *Mann test for trend*.

The Spearman coefficient of rank correlation (Section 11.3) is an alternative descriptive measure of association in a bivariate sample. Each set of observations is independently ranked from 1 to n, but the pairs are kept intact. The coefficient is given in (3.7) as

$$R = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

where D_i is the difference of the ranks of X_i and Y_i. If ties are present we use (3.19). Interpretation of the value of R is exactly the same as for T and the appropriate rejection regions are also in the same direction. For small samples the null distribution of R is given in Table M in a form similar to Table L. For large samples the rejection regions are simply $R \geqslant z_{\alpha/2}\sqrt{n-1}$ for positive dependence and $R \leqslant -z_{\alpha/2}\sqrt{n-1}$ for negative dependence. When R is used as a test for trend, it is frequently called the *Daniels' test for trend*. Applications of both of these statistics are illustrated in Example 6.1.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Alternative	*Rejection region*	*P value*
Positive dependence	$R \geqslant z_\alpha\sqrt{n-1}$	$P(Z \geqslant r/\sqrt{n-1})$
Negative dependence	$R \leqslant -z_\alpha\sqrt{n-1}$	$P(Z \leqslant r/\sqrt{n-1})$
Nonindependence	Both above with $z_{\alpha/2}$	2 (smaller of above)

Example 6.1 Two judges ranked nine teas on the basis of taste and full-bodied properties, with 1 indicating the highest ranking. Calculate the Kendall and Spearman measures of association, test the null hypothesis of independence, and find the appropriate one-tailed P value in each case, for the data shown below.

Tea	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Judge 1	1	5	9	7	4	6	8	2	3
Judge 2	4	3	6	8	2	7	9	1	5

Solution The first step in calculating Kendall's tau is to rearrange the data for Judge 1 in an array, keeping track of the corresponding rank of Judge 2 as shown below. Then the number of concordant pairs is counted as the number of Y ranks that are below and larger than each Y rank and then summed over all Y's; the number of discordant pairs is counted in the same manner but for ranks below and smaller.

Judge 1	*Judge 2*	*C*	*Q*	*D*	D^2
1	4	5	3	−3	9
2	1	7	0	1	1
3	5	4	2	−2	4
4	2	5	0	2	4
5	3	4	0	2	4
6	7	2	1	−1	1
7	8	1	1	−1	1
8	9	0	1	−1	1
9	6			3	9
		—	—	—	—
		28	8	0	34

We then calculate $T = 2(20)/9(8) = 0.556$. For the null hypothesis of independence the right-tailed P value from Table L is 0.022.

The last two columns above show that $\sum D_i^2 = 34$ and we compute $R = 1 - 6(34)/9(80) = 0.717$, which is larger than T as expected.

The right-tailed P value from Table M is $P = 0.018$ for the alternative of positive dependence.

At the time of this writing, MINITAB has no command for either Kendall's tau or Spearman's rho. However, we can use MINITAB to calculate Spearman's rho by using the rank command on the data (for Judges 1 and 2, respectively) and then calculating the Pearson product-moment correlation coefficient on these ranks. The result $R = 0.717$ agrees with ours. The MINITAB P value is for a Pearson correlation and does not apply for Spearman's rho.

The STATXACT solution gives the coefficients and the exact P values for a test of independence using both tau and rho, and all of these agree with ours. Note that the printout shows calculation of both τ_a and τ_b. These are equal because there are no ties in this example. The solution also shows τ_c and Somers' d, which apply for data in a contingency table and are not covered in this book. For Kendall's tau, STATXACT shows the asymptotic P-value based on the normal approximation $P(Z \geqslant 2.09)$ calculated from (2.30). For Spearman's rho, it shows the asymptotic P value based on the approximation given in (3.15) using Student's t distribution, $P(t \geqslant 2.72)$ with 7 degrees of freedom. The expressions they use for calculating the asymptotic standard errors and confidence interval estimates are not clear. The reader may verify, however, that they did not use our (2.36) because this gives an estimate of the variance of T which is negative in this example. As explained earlier, the estimate can be negative for n small, even though the exact value of the variance must be positive.

```
************************************
   MINITAB SOLUTION TO EXAMPLE 6.1
************************************

          C1    C2    C3    C4
           1     4     1     4
           5     3     5     3
           9     6     9     6
           7     8     7     8
           4     2     4     2
           6     7     6     7
           8     9     8     9
           2     1     2     1
           3     5     3     5

Correlations: C3, C4

Pearson correlation of C3 and C4 = 0.717
P-Value = 0.030
```


MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

```
*************************************
STATXACT SOLUTION TO EXAMPLE 6.1
*************************************

KENDALL'S TAU AND SOMER'S D RANK-ORDER CORRELATION COEFFICIENTS

Correlation Coefficient estimates based on 9 observations.

  Coefficient          Estimate      ASE1    95.00% Confidence Interval
  -----------          --------      ----    --------------------------
  Kendall's Tau          0.5556     0.1309   (     0.2989,      0.8122)
  Kendall's Tau_b        0.5556     0.1309   (     0.2989,      0.8122)
  Kendall's Tau_c        0.5556     0.1309   (     0.2989,      0.8122)
  Somers' D|row          0.5556     0.1309   (     0.2989,      0.8122)
  Somers' D|col          0.5556     0.1309   (     0.2989,      0.8122)
  Somers' D symm.        0.5556     0.1309   (     0.2989,      0.8122)

Asymptotic p-values for testing no association (Using Normal approximation):
        One-sided: Pr { Statistic .GE. Observed   } =     0.0000
        Two-sided: 2 * One-sided                    =     0.0000

Exact p-values for testing no association:
        One-sided: Pr {  Statistic  .GE.  Observed  }  =     0.0223
                   Pr {  Statistic  .EQ.  Observed  }  =     0.0099
        Two-sided: Pr { |Statistic| .GE. |Observed| }  =     0.0446

SPEARMAN'S CORRELATION TEST

Correlation Coefficient estimates based on 9 observations.

  Coefficient        Estimate        ASE1    95.00% Confidence Interval
  -----------        --------        ----    --------------------------
  Spearman's CC        0.7167      0.1061    (     0.5088,      0.9246)

Asymptotic p-values for testing no association (t-distribution with 7 df):
        One-sided: Pr { Statistic .GE. Observed   } =     0.0149
        Two-sided: 2 * One-sided                    =     0.0298

Exact p-values:
        One-sided: Pr {  Statistic  .GE.  Observed  }  =     0.0184
                   Pr {  Statistic  .EQ.  Observed  }  =     0.0029
        Two-sided: Pr { |Statistic| .GE. |Observed| }  =     0.0369
```

We use the data in Example 6.1 to illustrate how T can be interpreted as a coefficient of disarray, where Q, the number of discordant pairs, is the minimum number of interchanges in the Y ranks, one pair at a time, needed to convert them to the natural order. The X and Y ranks in this example are as follows.

X	1	2	3	4	5	6	7	8	9
Y	4	1	5	2	3	7	8	9	6

In the Y ranks, we first interchange the 4 and 1 to put 1 in the correct position. Then we interchange 2 and 5 to make 2 closer to its correct

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

position. Then we interchange 2 and 4. We keep proceeding in this way, working to get 3 in the correct position, and then 4, etc. The complete set of changes is as follows:

Y	1	4	5	2	3	7	8	9	6
	1	4	2	5	3	7	8	9	6
	1	2	4	5	3	7	8	9	6
	1	2	4	3	5	7	8	9	6
	1	2	3	4	5	7	8	9	6
	1	2	3	4	5	7	8	9	6
	1	2	3	4	5	7	8	6	9
	1	2	3	4	5	6	7	8	9

The total number of interchanges required to transform the Y ranks into the natural order by this systematic procedure is 8, and this is the value of Q, the total number of discordant pairs. We could make the transformation using more interchanges, of course, but more are not needed. It can be shown that $Q = 8$ is the minimum number of interchanges.

11.7 SUMMARY

In this chapter we have studied in detail the nonparametric coefficients that were proposed by Kendall and Spearman to measure association. Both coefficients can be computed for a sample from a bivariate distribution, a sample of pairs, when the data are numerical measurements or ranks indicating relative magnitudes. The absolute values of both coefficients range between zero and one, with increasing values indicating increasing degrees of association. The sign of the coefficient indicates the direction of the association, direct or inverse. The values of the coefficients are not directly comparable, however. We know that $|R| \geqslant |T|$ for any set of data, and in fact $|R|$ can be as much as 50 percent greater than $|T|$.

Both coefficients can be used to test the null hypothesis of independence between the variables. Even though the magnitudes of R and T are not directly comparable, the magnitudes of the P values based on them should be about the same, allowing for the fact that they are measuring association in different ways. The interpretation of T is easier than for R. T is the proportion of concordant pairs in the sample minus the proportion of discordant pairs. T can also be interpreted as a coefficient of disarray. The easiest interpretation of R is as the sample value of the Pearson product-moment correlation coefficient calculated using the ranks of the sample data.

An exact test of the null hypothesis of independence can be carried out using either T or R for small sample sizes. Generation of tables for exact P values was difficult initially, but now computers have the capacity for doing this for even moderate n. For intermediate and large sample sizes, the tests can be performed using large sample approximations. The distribution of T approaches the normal distribution much more rapidly than the distribution of R and hence approximate P values based on R are less reliable than those based on T.

Both T and R can be used when ties are present in either or both samples, and both have a correction for ties that improves the normal approximation. The correction with T always increases the value of T while the R correction always decreases the value of R, making the coefficients closer in magnitude.

If we reject the null hypothesis of independence by either T or R, we can conclude that there is some kind of dependence or "association" between the variables. But the kind of relationship or association that exists defies any verbal description in general. The existence of a relationship or significant association does not mean that the relationship is causal. The relationship may be due to several other factors, or to no factor at all. Care should always be taken in stating the results of an experiment that no causality is implied, either directly or indirectly.

Kendall's T is an unbiased estimator of a parameter τ in the bivariate population; τ represents the probability of concordance minus the probability of discordance. Concordance is not the same as correlation, although both represent a kind of association. Spearman's R is not an unbiased estimator of the population correlation ρ. It is an unbiased estimator of a parameter which is a function of τ and the grade correlation.

The tests of independence based on T and R can be considered nonparametric counterparts of the test that the Pearson product-moment correlation coefficient ρ is equal to zero in the bivariate normal distribution or that the regression coefficient β equals zero. The asymptotic relative efficiency of these tests relative to the parametric test based on the sample Pearson product-moment correlation coefficient is $9/\pi^2 = 0.912$ for normal distributions and one for the continuous uniform distribution.

Both T and R can be used to test for the existence of trend in a set of time-ordered observations. The test based on T is called the Mann test, and the test based on R is called the Daniels' test. Both of these tests are alternatives to the tests for randomness presented in Chapter 3.

PROBLEMS

11.1. A beauty contest has eight contestants. The two judges are each asked to rank the contestants in a preferential order of pulchritude. The results are shown in the table. Answer parts (*a*) and (*b*) using (i) the Kendall tau-coefficient procedures and (ii) the Spearman rank-correlation-coefficient procedures:

	Contestant							
Judge	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
1	2	1	3	5	4	8	7	6
2	1	2	4	5	7	6	8	3

(*a*) Calculate the measure of association.
(*b*) Test the null hypothesis that the judges ranked the contestants independently (use tables).
(*c*) Find a 95 percent confidence-interval estimate of τ.

11.2. Verify the result given in (4.9).

11.3. Two independent random samples of sizes m and n contain no ties. A set of $m+n$ paired observations can be derived from these data by arranging the combined samples in ascending order of magnitude and (*a*) assigning ranks, (*b*) assigning sample indicators. Show that Kendall's tau, calculated for these pairs without a correction for ties, is linearly related to the Mann-Whitney U statistic for these data, and find the relation if the sample indicators are (i) sample numbers 1 and 2, (ii) 1 for the first sample and 0 for the second sample as in the Z vector of Chapter 7.

11.4. Show that for the standardized bivariate normal distribution

$$\Phi(0,0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin \rho$$

11.5. The Census Bureau reported that Hispanics are expected to overtake blacks as the largest minority in the United States by the year 2030. Use *two different tests* to see whether there is a direct relationship between number of Hispanics and percent of state population for the nine states below.

State	*Hispanics (millions)*	*Percent of state population*
California	6.6	23
Texas	4.1	24
New York	2.1	12
Florida	1.5	12
Illinois	0.8	7
Arizona	0.6	18
New Jersey	0.6	8
New Mexico	0.5	35
Colorado	0.4	11

11.6. Company-financed expenditures in manufacturing on research and development (R&D) are currently about 2.7 percent of sales in Japan and 2.8 percent of sales in the United States. However, when these figures are looked at separately according to industry, the following data from Mansfield (1989) show some large differences.

Industry	*Japan*	*United States*
Food	0.8	0.4
Textiles	1.2	0.5
Paper	0.7	1.3
Chemicals	3.8	4.7
Petroleum	0.4	0.7
Rubber	2.9	2.2
Ferrous metals	1.9	0.5
Nonferrous metals	1.9	1.4
Metal products	1.6	1.3
Machinery	2.7	5.8
Electrical equipment	5.1	4.8
Motor vehicles	3.0	3.2
Other transport equipment	2.6	1.2
Instruments	4.5	9.0

(*a*) Use the signed-rank test to determine whether Japan spends a larger percentage than the United States on R&D.

(*b*) Determine whether there is a significant positive relationship between percentages spent by Japan and the United States (two different methods).

11.7. The *World Almanac and Book of Facts* published the following divorce rates per 1000 population in the United States. Determine whether these data show a positive trend using *four different* methods.

Year	*Divorce rate*
1945	3.5
1950	2.6
1955	2.3
1960	2.2
1965	2.5
1970	3.5
1975	4.8
1980	5.2
1985	5.0

11.8. For the time series data in Example 4.1 of Chapter 3, use the Mann test based on Spearman's rank correlation coefficient to see if the data show a positive trend.

11.9. Do Problem 11.8 using the Daniels' test based on Kendall's tau.

11.10. The rainfall measured by each of 12 gauges was recorded for 20 successive days. The average results for each day are as follows:

Day	*Rainfall*	*Day*	*Rainfall*
April 1	0.00	April 11	2.10
April 2	0.03	April 12	2.25
April 3	0.05	April 13	2.50
April 4	1.11	April 14	2.50
April 5	0.00	April 15	2.51
April 6	0.00	April 16	2.60
April 7	0.02	April 17	2.50
April 8	0.06	April 18	2.45
April 9	1.15	April 19	0.02
April 10	2.00	April 20	0.00

Use an appropriate test to determine whether these data exhibit some sort of pattern. Find the P value:

(a) Using tests based on runs with both the exact distribution and the normal approximation.
(b) Using other tests that you may think are appropriate.
(c) Compare and interpret the results of (a) and (b).

11.11 A company has administered a screening aptitude test to 20 new employees over a two-year period. The record of scores and date on which the person was hired are shown below.

1/4/01	75	9/21/01	72	12/9/01	81	5/10/02	91
3/9/01	74	10/4/01	77	1/22/02	93	7/17/02	95
6/3/01	71	10/9/01	76	1/26/02	82	9/12/02	90
6/15/01	76	11/1/01	78	3/21/02	84	10/4/02	92
8/4/01	98	12/5/01	80	4/6/02	89	12/6/02	93

Assuming that these test scores are the primary criterion for hiring, do you think that over this time period the screening procedure has changed, or the personnel agent has changed, or supply has changed, or what? Base your answer on an appropriate nonparametric procedure (there are several appropriate methods).

11.12. Ten randomly chosen male college students are used in an experiment to investigate the claim that physical strength is decreased by fatigue. Describe the relationship for the data below, using several methods of analysis.

Minutes between rest periods	*Pounds lifted per minute*
5.5	350
9.6	230
2.4	540
4.4	390
0.5	910
7.9	220
2.0	680
3.3	590
13.1	90
4.2	520

11.13. Given a single series of time-ordered ordinal observations over several years, name some nonparametric procedures that could be used and how in order to detect a long-term positive trend. Name as many as you can think of.

11.14. Six randomly selected mice are studied over time and scored on an ordinal basis for intelligence and social dominance. The data are as follows:

Mouse	*Intelligence*	*Social dominance*
1	45	63
2	26	0
3	20	16
4	40	91
5	36	25
6	23	2

(a) Find the coefficient of rank correlation.
(b) Find the appropriate one-tailed P value for your result in (a).
(c) Find the Kendall tau coefficient.
(d) Find the appropriate one-tailed P value for your result in (c).

11.15. A board of marketing executives ranked 10 similar products, and an "independent" group of male consumers also ranked the products. Use two different nonparametric procedures to describe the correlation between rankings and find a one-tailed P value in each case. State the hypothesis and alternative and all assumptions. Compare and contrast the procedures.

Product	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Executive ranks	9	4	3	7	2	1	5	8	10	6
Independent male ranks	7	6	5	9	2	3	8	5	10	1

11.16. Derive the null distribution of both Kendall's tau statistic and Spearman's rho for $n = 3$ assuming no ties.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

11.17. A scout for a professional baseball team ranks nine players separately in terms of speed and power hitting, as shown below.

Player	*Speed ranking*	*Power-hitting ranking*
A	3	1
B	1	3
C	5	4
D	6	2
E	2	6
F	7	8
G	8	9
H	4	5
I	9	7

(*a*) Find the rank correlation coefficient and the appropriate one-tailed P value.
(*b*) Find the Kendall tau coefficient and the appropriate one-tailed P value.

11.18. Twenty-three subjects are asked to give their attitude toward elementary school integration and their number of years of schooling completed. The data are shown below.

Number of years of school completed at the time	*Attitude toward elementary school integration*			
	Strongly disagree	*Moderately disagree*	*Moderately agree*	*Strongly agree*
0–6	5	9	12	16
7–9	4	10	13	18
10–12 or G.E.D.	10	7	9	12
Some college	12	12	12	19
College degree (4 yr)	3	12	16	14
Some Graduate		10	15	
Graduate degree			14	

As a measure of the association between attitude and number of years of schooling completed:

(*a*) Compute Kendall's tau with correction for ties.
(*b*) Compute Spearman's R with correction for ties.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

12

Measures of Association in Multiple Classifications

12.1 INTRODUCTION

Suppose we have a set of data presented in the form of a complete two-way layout of I rows and J columns, with one entry in each of the IJ cells. In the sampling situation of Chapter 10, if the independent samples drawn from each of I univariate populations were all of the same size J, we would have a complete layout of IJ cells. However, this would be called a one-way layout since only one factor is involved, the populations. Under the null hypothesis of identical populations, the data can be considered a single random sample of size IJ from the common population. The parallel to this problem in classical statistics is the one-way analysis of variance. In this chapter we shall study some nonparametric analogs of the two-way analysis-of-variance problem, all parallel in the sense that the data are presented in the form of a two-way layout which cannot be considered a single random sample because of certain relationships among elements.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Let us first review the techniques of the analysis-of-variance approach to testing the null hypothesis that the column effects are all the same. The model is usually written

$$X_{ij} = \mu + \beta_i + \theta_j + E_{ij} \qquad \text{for } i = 1, 2, \ldots, I \quad \text{and} \quad j = 1, 2, \ldots, J$$

The β_i and θ_j are known as the row and column effects, respectively. In the normal-theory model, the errors E_{ij} are independent, normally distributed random variables with mean zero and variance σ_E^2. The test statistic for the null hypothesis of equal column effects or, equivalently,

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_J$$

is the ratio

$$\frac{(I-1)I\sum_{j=1}^{J}(\bar{x}_j - \bar{x})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2}$$

where

$$\bar{x}_i = \sum_{j=1}^{J}\frac{x_{ij}}{J} \qquad \bar{x}_j = \sum_{i=1}^{I}\frac{x_{ij}}{I} \qquad \bar{x} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{x_{ij}}{IJ}$$

If all the assumptions of the model are met, this test statistic has the F distribution with $J-1$ and $(I-1)(J-1)$ degrees of freedom.

The first two parallels of this design which we shall consider are the k-related or k-matched sample problems. The matching can arise in two different ways, but both are somewhat analogous to the randomized-block design of a two-way layout. In this design, IJ experimental units are grouped into I blocks, each containing J units. A set of J treatments is assigned at random to the units within each block in such a way that all J assignments are equally likely, and the assignments in different blocks are independent. The scheme of grouping into blocks is important, since the purpose of such a design is to minimize the differences between units in the same block. If the design is successful, an estimate of experimental error can be obtained which is not inflated by differences between blocks. This model is often appropriate in agricultural field experimentation since the effects of a possible fertility gradient can be reduced. Dividing the field into I blocks, the plots within each block can be kept in close proximity. Any differences between plots within the same block can be attributed to differences between treatments and the block effect can be eliminated from the estimate of experimental error.

The first related-sample problem arises where IJ subjects are grouped into I blocks each containing J-matched subjects, and within each block J treatments are assigned randomly to the matched subjects. The effects of the treatments are observed, and we let X_{ij} denote the observation in block i of treatment number $j, i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$. Since the observations in different blocks are independent, the collection of entries in column number j are independent. In order to determine whether the treatment (column) effects are all the same, the analysis-of-variance test is appropriate if the requisite assumptions are justified. If the observations in each row $X_{i1}, X_{i2}, \ldots, X_{iJ}$ are replaced by their ranking within that row, a nonparametric test involving the column sums of this $I \times J$ table, called *Friedman's two-way analysis of variance by ranks*, can be used to test the same hypothesis. This is a k-related sample problem when $J = k$. This design is sometimes called a *balanced complete block design* and also a *repeated measures design*. The null hypothesis is that the treatment effects are all equal or

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_J$$

and the alternative for the Friedman test is

$$H_1: \theta_i \neq \theta_j \quad \text{for at least one } i \neq j$$

A related nonparametric test for the k-related sample problem is called Page's test for ordered alternatives. The null hypothesis is the same as above but the alternative specifies the treatment effects as occurring in a specific order, as for example,

$$H_1: \theta_1 < \theta_2 < \cdots < \theta_J$$

For each of these problems the location model is that the respective cdf's are $F(x - \theta_i - \beta_j)$.

Another related-sample problem arises by considering a single group of J subjects, each of which is observed under I different conditions. The matching here is by condition rather than subject, and the observation X_{ij} denotes the effect of condition i on subject number $j, i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$. We have here a random sample of size J from an I-variate population. Under the null hypothesis that the I variates are independent, the expected sum of the I observations on subject number j is the same for all $j = 1, 2, \ldots, J$. In order to determine whether the column effects are all the same, the analysis-of-variance test may be appropriate. Testing the independence of the I variates involves a comparison of J column totals, so that in a sense the roles of treatments and blocks have been reversed in terms of which factor is of interest. This is a k-related sample problem when $I = k$. If the observations in each row are ranked as before, Friedman's two-way analysis of variance will provide a nonparametric test of independence of the k variates. Thus, in order to effect consistency of results as opposed to consistency of sampling situations, the presentation here in both cases will be for a table containing k rows and n columns, where each row is a set of positive integer ranks.

In this second related-sample problem, particularly if the null hypothesis of the independence of the k variates is rejected, a measure of the association between the k variates would be desirable. In fact, this sampling situation is the direct extension of the paired-sample problem of Chapter 11 to the k-related sample case. A measure of the overall agreement between the k sets of rankings, called *Kendall's coefficient of concordance*, can be determined. This statistic can also be used to test the null hypothesis of independence, but the test is equivalent to Friedman's test for n treatments and k blocks. An analogous measure of concordance will be found for k sets of incomplete rankings, which relate to the balanced incomplete-blocks design. Another topic to be treated briefly is a nonparametric approach to finding a measure of partial correlation or correlation between two variables when a third is held constant when there are three complete sets of rankings of n objects.

12.2 FRIEDMAN'S TWO-WAY ANALYSIS OF VARIANCE BY RANKS IN A $k \times n$ TABLE AND MULTIPLE COMPARISONS

As suggested in Section 12.1, in the first related-sample problem we have data presented in the form of a two-way layout of k rows and n columns. The rows indicate block, subject, or sample numbers, and the columns are treatment numbers. The observations in different rows are independent, but the columns are not because of some unit of association. In order to avoid making the assumptions requisite for the usual analysis of variance test that the n treatments are the same, Friedman (1937, 1940) suggested replacing each treatment observation within the ith block by a number from the set $\{1, 2, \ldots, n\}$ which represents that treatment's magnitude relative to the other observations in the same block. We denote the ranked observations by R_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$, so that R_{ij} is the rank of treatment number j when observed in block number i. Then $R_{i1}, R_{i2}, \ldots, R_{in}$ is a permutation of the first n integers, and $R_{1j}, R_{2j}, \ldots, R_{kj}$ is the set of

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

ranks given to treatment number j in all blocks. We represent the data in tabular form as follows:

$$
\begin{array}{llcccccl}
 & & \multicolumn{4}{c}{\textit{Treatments}} & & \\
 & & 1 & 2 & \cdots & n & & \textit{Row totals} \\
 & 1 & R_{11} & R_{12} & \cdots & R_{1n} & & n(n+1)/2 \\
 & 2 & R_{21} & R_{22} & \cdots & R_{2n} & & n(n+1)/2 \\
 & . & & & & & & . \\
\textit{Blocks} & . & \multicolumn{4}{c}{\cdots\cdots\cdots\cdots} & & . \\
 & . & & & & & & . \\
 & k & R_{k1} & R_{k2} & \cdots & R_{kn} & & n(n+1)/2 \\
\multicolumn{2}{l}{\textit{Column totals}} & R_1 & R_2 & \cdots & R_n & & kn(n+1)/2
\end{array}
\tag{2.1}
$$

The row totals are of course constant, but the column totals are affected by differences between treatments. If the treatment effects are all the same, each expected column total is the same and equals the average column total $k(n+1)/2$. The sum of deviations of observed column totals around this mean is zero, but the sum of squares of these deviations will be indicative of the differences in treatment effects. Therefore we shall consider the sampling distribution of the random variable.

$$S = \sum_{j=1}^{n}\left[R_j - \frac{k(n+1)}{2}\right]^2 = \sum_{j=1}^{n}\left[\sum\nolimits_{i=1}^{k}\left(R_{ij} - \frac{n+1}{2}\right)\right]^2 \tag{2.2}$$

under the null hypothesis of no difference between the n treatment effects, that is,

$$H_0\colon \theta_1 = \theta_2 = \cdots = \theta_n$$

For this null case, in the ith block the ranks are assigned completely at random, and each row in the two-way layout constitutes a random permutation of the first n integers if there are no ties. There are then a total of $(n!)^k$ distinguishable sets of entries in the $k \times n$ table, and each is equally likely. These possibilities can be enumerated and the value of S calculated for each. The probability distribution of S then is

$$f_S(s) = \frac{u_s}{(n!)^k}$$

where u_s is the number of those assignments which yield s as the sum of squares of column total deviations. A systematic method of

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

generating the values of u_s for n, k from the values of u_s for $n, k-1$ can be employed (see Kendall and Gibbons, 1990, pp. 150–151). A table of the distribution of S is given here in Table N of the Appendix for $n=3$, $k \leqslant 8$ and $n=4$, $k \leqslant 4$. More extensive tables for the distribution of Q, a linear function of S to be defined later in (2.8), are given in Owen (1962) for $n=3$, $k \leqslant 15$ and $n=4$, $k \leqslant 8$. Other tables are given in Michaelis (1971), Quade (1972), and Odeh (1977) that cover the cases up to $k=6$, $n=6$. However, the calculations are considerable even using the systematic approach. Therefore, outside the range of existing tables, an approximation to the null distribution is generally used for tests of significance.

Using the symbol μ to denote $(n+1)/2$, (2.2) can be written as

$$\begin{aligned} S &= \sum_{j=1}^{n}\sum_{i=1}^{k}(R_{ij}-\mu)^2 + 2\sum_{j=1}^{n}\sum\sum_{1\leqslant i<p\leqslant k}(R_{ij}-\mu)(R_{pj}-\mu) \\ &= k\sum_{j=1}^{n}(j-\mu)^2 + 2U \\ &= \frac{kn(n^2-1)}{12} + 2U \end{aligned} \tag{2.3}$$

The moments of S then are determined by the moments of U, which can be found using the following relations from (3.2), (3.3), and (3.10) of Chapter 11:

$$E(R_{ij}) = \frac{n+1}{2} \qquad \operatorname{var}(R_{ij}) = \frac{n^2-1}{12}$$

$$\operatorname{cov}(R_{ij}, R_{iq}) = -\frac{n+1}{12}$$

Furthermore, by the design assumptions, observations in different rows are independent, so that for all $i \neq p$ the expected value of a product of functions of R_{ij} and R_{pq} is the product of the expected values and $\operatorname{cov}(R_{ij}, R_{pq}) = 0$. Then

$$E(U) = n\binom{k}{2}\operatorname{cov}(R_{ij}, R_{pj}) = 0$$

so that $\operatorname{var}(U) = E(U^2)$, where

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$U^2 = \sum_{j=1}^{n} \sum\sum_{1 \leqslant i<p \leqslant k} (R_{ij} - \mu)^2 (R_{pj} - \mu)^2$$
$$+ 2 \sum\sum_{1 \leqslant j<q \leqslant n} \sum\sum_{1 \leqslant i<p \leqslant k} \sum\sum_{1 \leqslant r<s<k} (R_{ij} - \mu)(R_{pj} - \mu)(R_{rq} - \mu)(R_{sq} - \mu) \tag{2.4}$$

Since R_{ij} and R_{pq} are independent whenever $i \neq p$, we have

$$E(U^2) = \sum_{j=1}^{n} \sum\sum_{1 \leqslant i<p \leqslant k} \operatorname{var}(R_{ij})\operatorname{var}(R_{pj})$$
$$+ 2 \sum\sum_{1 \leqslant j<q \leqslant n} \binom{k}{2} \operatorname{cov}(R_{ij}, R_{iq}) \operatorname{cov}(R_{pj}, R_{pq}) \tag{2.5}$$

$$E(U^2) = n\binom{k}{2}\frac{(n^2-1)^2}{144} + 2\binom{n}{2}\binom{k}{2}\frac{(n+1)^2}{144}$$
$$= n^2\binom{k}{2}(n+1)^2\frac{(n-1)}{144} \tag{2.6}$$

Substituting these results back in (2.3), we find

$$E(S) = \frac{kn(n^2-1)}{12} \qquad \operatorname{var}(S) = \frac{n^2k(k-1)(n-1)(n+1)^2}{72} \tag{2.7}$$

A linear function of the random variables defined as

$$Q = \frac{12S}{kn(n+1)} = \frac{12\sum_{j=1}^{n} R_j^2}{kn(n+1)} - 3k(n+1) \tag{2.8}$$

has moments $E(Q) = n-1, \operatorname{var}(Q) = 2(n-1)(k-1)/k \approx 2(n-1)$, which are the first two moments of a chi-square distribution with $n-1$ degrees of freedom. The higher moments of Q are also closely approximated by corresponding higher moments of the chi-square for k large. For all practical purposes then, Q can be treated as a chi-square variable with $n-1$ degrees of freedom. Numerical comparisons have shown this to be a good approximation as long as $k > 7$. The rejection region for a test of equal treatment effects against the alternative that the effects are not all equal with significance level approximately α is

$$Q \in R \qquad \text{for } Q \geqslant \chi^2_{n-1,\alpha}$$

A test based on S or Q is called *Friedman's test*.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

From classical statistics, we are accustomed to thinking of an analysis-of-variance test statistic as the ratio of two estimated variances or mean squares of deviations. The total sum of squares of deviations of all nk ranks around the average rank is

$$s_t = \sum_{i=1}^{k}\sum_{j=1}^{n}(r_{ij}-\bar{r})^2 = k\sum_{j=1}^{n}\left(j-\frac{n+1}{2}\right)^2 = kn\frac{n^2-1}{12}$$

and thus we could write Friedman's test statistic in (2.8) as

$$Q = \frac{(n-1)S}{s_t}$$

Even though s_t is a constant, as in classical analysis-of-variance problems, it can be partitioned into a sum of squares of deviations between columns plus a residual sum of squares. Denoting the grand mean and column means respectively by

$$\bar{r} = \sum_{i=1}^{k}\sum_{j=1}^{n}\frac{r_{ij}}{nk} = \frac{n+1}{2} \qquad \bar{r}_j = \frac{r_j}{k} = \sum_{i=1}^{k}\frac{r_{ij}}{k}$$

we have

$$\begin{aligned}
s_t &= \sum_{i=1}^{k}\sum_{j=1}^{n}(r_{ij}-\bar{r})^{\hat{2}} = \sum_{i=1}^{k}\sum_{j=1}^{n}(r_{ij}-\bar{r}_j+\bar{r}_j-\bar{r})^2\\
&= \sum_{i=1}^{k}\sum_{j=1}^{n}(r_{ij}-\bar{r}_j)^2 + k\sum_{j=1}^{n}(\bar{r}_j-\bar{r})^2 + 2\sum_{j=1}^{n}(\bar{r}_j-\bar{r})\sum_{i=1}^{k}(r_{ij}-\bar{r}_j)\\
&= \sum_{i=1}^{k}\sum_{j=1}^{n}(r_{ij}-\bar{r}_j)^2 + \sum_{j=1}^{n}\frac{\{r_j-[k(n+1)/2]\}^2}{k}
\end{aligned}$$

or

$$s_t = \sum_{i=1}^{k}\sum_{j=1}^{n}(r_{ij}-\bar{r}_j)^2 + \frac{s}{k} = kn\frac{n^2-1}{12} \tag{2.9}$$

An analogy to the classical analysis-of-variance table is given in Table 2.1.

The usual statistic for equal column effects is the ratio of the column and residual mean squares, or

$$\frac{(k-1)S}{ks_t - S} \tag{2.10}$$

If the distributions are normal with equal variances, the null distribution of the statistic in (2.10) is Snedecor's F with $(n-1)$ and $(n-1)(k-1)$ degrees of freedom.

APPLICATIONS

Friedman's two-way analysis of variance by ranks is appropriate for the null hypothesis of equal treatment effects

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_n$$

for data on n treatments applied in k blocks. We note that the word treatment effect is used in a very general way and may not refer to a real treatment. It may refer to the effect of a condition or characteristic such as income level or race. The first step is to rank the data in each block from 1 to n. These ranks are summed for each column to obtain $R_1, R_2, \ldots, R_n$. One form of the test statistic is S, the sum of squares of deviations of these totals from their mean, given in (2.2) but simplified here for calculation to

$$S = \sum_{j=1}^{n} R_j^2 - \frac{k^2 n(n+1)^2}{4} \tag{2.11}$$

The null distribution of S is given in Table N for $n = 3, k \leqslant 8$ and $n = 4, k \leqslant 4$ as right-tail probabilities since H_0 should be rejected for S large. For other n, k we can use Table B since the asymptotic distribution of Q given in (2.8) is chi square with $n-1$ degrees of freedom.

If ties are present to the extent t, we use midranks. The test statistic that incorporates the correction for ties is

Table 2.1 Analysis of variance table

Source of variation	*Degrees of freedom*	*Sum of squares*
Between columns	$n-1$	s/k
Between rows	$k-1$	0[a]
Residual	$(n-1)(k-1)$	$s_t - s/k$
Total	$nk-1$	s_t

[a] There is no variation between rows here since the row sums are all equal.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

$$Q = \frac{12(n-1)S}{kn(n^2-1) - \sum\sum t(t^2-1)} \tag{2.12}$$

where the double sum is extended over all sets of t tied ranks in each of the k blocks. This result will be derived in Section 12.4.

If the null hypothesis of equal treatment effects is rejected, we may want to determine which pairs of treatments differ significantly in effects and in which direction. Then we can use a multiple comparisons procedure to compare the $n(n-1)/2$ pairs of treatments, as we did in Section 10.4 for the one-way analysis-of-variance test for equal medians.

The procedure is to declare that treatments i and j are significantly different in effect if

$$|R_i - R_j| \geqslant z^* \sqrt{kn(n+1)/6} \tag{2.13}$$

where z^* is found as in Section 10.4 as the negative of the $[\alpha/n(n-1)]$th quantile of the standard normal distribution. As before, α is generally chosen to be larger than in the typical hypothesis testing situation, as around 0.15 or 0.20, because so many comparisons are being made.

Example 2.1 An important factor in raising small children is to develop their ability to ask questions, especially in groups so that they will have this skill when they start school. A study of group size and number of questions asked by preprimary children in a classroom atmosphere with a familiar person as teacher consists of dividing 46 children randomly into four mutually exclusive groups of sizes 24, 12, 6, and 4. The total number of questions asked by all members of each group is recorded for 30 minutes on each of eight different days. For the data shown in Table 2.2, test the null hypothesis that the effect of the group size is the same in terms of total number of questions asked.

Solution The days serve as blocks and the group sizes are the treatments so that $n = 4, k = 8$. The null hypothesis is equal treatment effects or H_0: $\theta_1 = \theta_2 = \theta_3 = \theta_4$. The first step is to rank the observations for each day from 1 to 4, using midranks for the few ties, and sum the columns to find the R_j, as shown in Table 2.3.

We calculate the sum of squares from (2.11) as

$$S = (8^2 + 17^2 + 26.5^2 + 28.5^2) - \frac{8^2(4)(5)^2}{4} = 267.5$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 2.2 Data for Example 2.1

	Group size			
Day	*24*	*12*	*6*	*4*
1	14	23	26	30
2	19	25	25	33
3	17	22	29	28
4	17	21	28	27
5	16	24	28	32
6	15	23	27	36
7	18	26	27	26
8	16	22	30	32

and then

$$Q = \frac{12(267.5)}{8(4)(5)} = 20.1$$

from (2.8) with 3 degrees of freedom. Table B of the Appendix shows that $P < 0.001$, so we reject the null hypothesis. It appears that the larger the group size, the fewer the questions asked.

Notice that there are two sets of ties, occurring on days 1 and 7, and each is of extent 2. Hence $\sum\sum t(t^2 - 1) = 12$ and the corrected test statistic from (2.12) is $Q = 20.58$. The P value is unchanged.

Since the difference between the n treatment effects has been found to be significant, we can use the multiple comparisons procedure to determine which pairs of treatments differ significantly. With $\alpha = 0.15$ say, $k = 8, n = 4$, we have $z^* = 2.241$ and the right-hand side of (2.13) is 11.572. The groups that differ significantly are sizes 6 and 24, and sizes 4 and 24.

Table 2.3 Ranks of data for Example 2.1

	Group size			
Day	*24*	*12*	*6*	*4*
1	1	2	3	4
2	1	2.5	2.5	4
3	1	2	4	3
4	1	2	4	3
5	1	2	3	4
6	1	2	3	4
7	1	2.5	4	2.5
8	1	2	3	4
Total	8	17	26.5	28.5

The computer solutions to this example are shown below from the MINITAB and STATXACT packages. We note that the correction for ties was incorporated to calculate Q in STATXACT while MINITAB gives the answer both with and without the correction. The exact P value in STATXACT is based on the randomization distribution or permutation distribution of the test statistic.

```
****************************************
   MINITAB SOLUTION TO EXAMPLE 2.1
****************************************
```

1	1	14
1	2	19
1	3	17
1	4	17
1	5	16
1	6	15
1	7	18
1	8	16
2	1	23
2	2	25
2	3	22
2	4	21
2	5	24
2	6	23
2	7	26
2	8	22
3	1	26
3	2	25
3	3	29
3	4	28
3	5	28
3	6	27
3	7	27
3	8	30
4	1	30
4	2	33
4	3	28
4	4	27
4	5	32
4	6	36
4	7	26
4	8	32

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

```
**************************************************
   MINITAB SOLUTION TO EXAMPLE 2.1 CONTINUED
**************************************************

Friedman Test: C3 versus C1,C2

Friedman test for C3 by C1 blocked by C2

S = 20.06 DF = 3 P = 0.000
S - 20.58 DF - 3 P - 0.000 (adjusted for ties)

                     Est    Sum of
C1           N    Median    Ranks
1            8    16.125      8.0
2            8    23.500     17.0
3            8    27.875     26.5
4            8    31.000     28.5

Grand median = 24.625


        ****************************************
          STATXACT SOLUTION TO EXAMPLE 2.1
        ****************************************

FRIEDMAN TEST
[That 4 treatments have identical effects in 8 informative blocks

Statistic based on the observed 4 by 8 Two-way layout (x):
   FR(x) :Friedman Statistic = 20.58


Asymptotic p-value: (based on Chi-square distribution with 3 df)
   Pr {FR(X).GE.   20.58}-   0.0001


Exact p-value and point probability:
   Pr{FR(X).GE.    20.58}=   0.0000
   Pr{FR(X).EQ.    20.58}-   0.0000
```

12.3 PAGE'S TEST FOR ORDERED ALTERNATIVES

The alternative hypothesis for Friedman's two-way analysis of variance by ranks in a $k \times n$ table described in Section 12.2 is that the treatment effects are not all the same, a two-sided alternative. Now suppose that we want a one-sided alternative or an ordered alternative that the treatment effects θ_i occur in a specified order, e.g.,

$$H_1: \theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_n$$

with at least one inequality strict. Page (1963) suggested a test based on a weighted sum of the column totals

$$L = \sum_{j=1}^{n} Y_j R_j \tag{3.1}$$

where the weight Y_j is the hypothetical ranking of the jth treatment, predicted from prior considerations. The null hypothesis should be rejected in favor of this ordered alternative for large values of L. Tables of exact critical values of L are given in Page (1963) for levels 0.001, 0.01, and 0.05 and reproduced here as Table Q in the Appendix. For large values of k and n, the statistic (with a continuity correction)

$$Z = \frac{12(L - 0.5) - 3kn(n+1)^2}{n(n+1)\sqrt{k(n-1)}} \tag{3.2}$$

is approximately standard normal and the appropriate rejection region is right tail.

The test based on L can be shown to be related to the average of the rank correlation coefficients between each ranking and the ranking predicted by the alternative. This relationship is

$$r_{av} = \frac{12L}{k(n^3 - n)} - \frac{3(n+1)}{n-1}$$

The Page test can also be used in the situation of Section 12.4 where we have k sets of rankings of n objects and the alternative states an a priori ranking of the objects.

Example 3.1 This numerical example is based on one used by Page (1963) to illustrate his proposed procedure. The research hypothesis is that speed of learning is related to the similarity of practice

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

material used in pretraining sessions to the test criterion learning material. Group A used practice material most similar to that of criterion learning, followed by groups B and C in that order, and group D had no pretraining. Therefore the predicated ranking from best to worst is A, B, C, D, or $\theta_D < \theta_C < \theta_B < \theta_A$, where rank 1 is given to the least rapid learning. Note that Page's article uses 1 to denote most rapid, whereas we use 1 to denote least rapid. Six different classes divided into these four groups gave the rankings shown in Table 3.1.

We use (3.1) to compute $L = 168$. The critical value from Table Q with $n = 4, k = 6$ and $\alpha = 0.05$ is 163, so we reject the null hypothesis of equal treatment effects in favor of the ordered alternative. Using the normal approximation with a continuity correction and $\alpha = 0.05$, we reject when

$$L \geqslant \frac{kn(n+1)^2}{4} + 0.5 + \frac{n(n+1)\sqrt{k(n-1)}}{12} z_\alpha$$

which equals 162.13 for our example ($z_{0.05} = 1.645$). Again, we reject the null hypothesis using the normal approximation.

The computer solution to this example is shown below with an output from STATXACT; the value of the statistic 168 agrees with ours and both the exact and the approximate P values suggest rejecting the null hypothesis, which agree with our conclusions. The reader can

Table 3.1 Data for Example 3.1

	Treatments			
Classes	*A*	*B*	*C*	*D*
1	3	4	2	1
2	4	2	1	3
3	4	2	3	1
4	4	1	3	2
5	2	4	3	1
6	4	3	1	2
	—	—	—	—
R_j	21	16	13	10
Y_j	4	3	2	1

verify that STATXACT does not use a continuity correction to calculate the approximate P value.

```
              *****************************************
                STATXACT SOLUTION TO EXAMPLE 3.1
              *****************************************
PAGE TEST

[ That 4 treatments have identical effects in each of 6 blocks]

Statistic based on the observed 4 by 6 two-way layout(x) :

     Mean        Std-dev   Observed(PA(x))  Standardized(PA*(x))
    150.0         7.071          168.0                2.546

Asymptotic p-value:
    One-sided: Pr { PA*(X) .GE.        2.546 }    =  0.0055
    Two-sided: 2 * One-sided                      =  0.0109

Exact p-value:
   One-sided: Pr {   PA*(X)  .GE.        2.546 }  =   0.0053
   Pr    {  PA*(X)     .EQ.     2.546 }          =     0.0021
   Two-sided: Pr  { |PA*(X)| .GE.   2.546 }      =     0.0106
```

Example 3.2 In light of our conjecture in Example 2.1, it will be instructive to repeat the data analysis for the alternative

$$H_1\colon \theta_{24} \leqslant \theta_{12} \leqslant \theta_6 \leqslant \theta_4$$

where θ indicates the effect of the group size on asking questions and the subscript indicates the size of the group. The test statistic for these data is

$$L = 1(8) + 2(17) + 3(26.5) + 4(28.5) = 235.5$$

and the P value from Table Q is less than 0.001. Our previous conjecture that the larger the group size, the fewer questions are asked does appear to be correct.

The computer solution to this example is shown below with an output from STATXACT. Our hand calculations and conclusions agree with those obtained from the output.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

```
        ****************************************
         STATXACT SOLUTION TO EXAMPLE 3.2
        ****************************************

PAGE TEST
[ That 4 treatments have identical effects in each of 6 blocks]

Statistic based on the observed 4 by 6 two-way layout(x) :

   Mean       Std-dev   Observed(PA(x))  Standardized(PA*(x))
  200.0       8.165        235.5               4.348

Asymptotic p-value:
    One-sided: Pr { PA*(X) .GE.       4.348 }  =      0.0000
    Two-sided: 2 * One-sided                   =      0.0000

Exact p-value:
   one-sided: Pr {  PA*(X)  .GE.      4.348 }  =      0.0000
              Pr {  PA*(X)  .EQ.      4.348 }  =      0.0000
   Two-sided: Pr {|PA*(X)|  .GE.      4.348 }  =      0.0000
```

A different test for ordered alternatives was proposed by Jonckheere (1954), but the Page test is easier to use.

12.4 THE COEFFICIENT OF CONCORDANCE FOR *k* SETS OF RANKINGS OF *n* OBJECTS

The second k-related sample problem mentioned in Section 12.1 involves k sets of rankings of n subjects, where we are interested both in testing the hypothesis that the k sets are independent and in finding a measure of the relationship between rankings. In the common parlance of this type of sample problem, the k conditions are called *observers*, each of whom is presented with the same set of n objects to be ranked. The measure of relationship then will describe the agreement or concordance between observers in their judgments on the n objects.

Since the situation here is an extension of the paired-sample problem of Chapter 11, one possibility for a measure of agreement is to select one of the measures for paired samples and apply it to each of the $\binom{k}{2}$ sets of pairs of rankings of n objects. However, if $\binom{k}{2}$ tests of the null hypothesis of independence are then made using the sampling distribution appropriate for the measure employed, the tests are not

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

independent and the overall probability of a type I error is difficult to determine but necessarily increased. Such a method of hypothesis testing is then statistically undesirable. We need a single measure of overall association which will provide a single test statistic designed to detect overall dependence between samples with a specified significance level. If we could somehow combine measures obtained for each of the $\binom{k}{2}$ pairs, this would provide a single coefficient of overall association which can be used to test the null hypothesis of independence or no association between rankings if its sampling distribution can be determined.

The *coefficient of concordance* is such an approach to the problem of relationship between k sets of rankings. It is a linear function of the average of the coefficients of rank correlation for all pairs of rankings, as will be shown later in this section. However, the rationale of the measure will be developed independently of the procedures of the last chapter so that the analogy to analysis-of-variance techniques will be more apparent.

For the purpose of this parallel, then, we visualize the data as presented in the form of a two-way layout of dimension $k \times n$ as in (2.1), with row and column labels now designating observers and objects instead of blocks and treatments. The table entries R_{ij} denote the rank given by the ith observer to the jth object. Then the ith row is a permutation of the numbers $1, 2, \ldots, n$, and the jth column is the collection of ranks given to object number j by all observers. The ranks in each column are then indicative of the agreement between observers, since if the jth object has the same preference relative to all other objects in the opinion of each of the k observers, all ranks in the jth column will be identical. If this is true for every column, the observers agree perfectly and the respective column totals $(R_1, R_2, \ldots, R_n)$ will be some permutation of the numbers

$$1k, 2k, 3k, \ldots, nk$$

Since the average column total is $k(n+1)/2$, for perfect agreement between rankings the sum of squares of deviations of column totals from the average column total will be a constant

$$\sum_{j=1}^{n}\left[jk - \frac{k(n+1)}{2}\right]^2 = k^2 \sum_{j=1}^{n}\left(j - \frac{n+1}{2}\right)^2 = k^2 n \frac{n^2-1}{12} \tag{4.1}$$

The actual observed sum of squares of these deviations is

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

$$S = \sum_{j=1}^{n} \left[R_j - \frac{k(n+1)}{2} \right]^2 \tag{4.2}$$

We found in (2.9) that

$$ks_t = \frac{k^2 n(n^2-1)}{12} = s + k \sum_{i=1}^{k} \sum_{j=1}^{k} (r_{ij} - \bar{r}_j)^2 \tag{4.3}$$

where s_t is the total sum of squares of deviations of all ranks around the average rank. In terms of this situation, however, we see from (4.1) that ks_t is the sum of squares of column total deviations when there is perfect agreement. Therefore the value of S for any set of k rankings ranges between zero and $k^2 n(n^2-1)/12$, with the maximum value attained when $r_j = jk$ for all j, that is, when there is perfect agreement, and the minimum value attained when $r_j = k(n+1)/2$ for all j, that is, when each observer's rankings are assigned completely at random so that there is no agreement between observers. If the observers are called samples, no agreement between observers is equivalent to independence of the k samples.

The ratio of S to its maximum value

$$W = \frac{S}{ks_t} = \frac{12S}{k^2 n(n^2-1)} \tag{4.4}$$

therefore provides a measure of agreement between observers, or concordance between sample rankings, or dependence of the samples. This measure is called *Kendall's coefficient of concordance*. It ranges between 0 and 1, with 1 designating perfect agreement or concordance and 0 indicating no agreement or independence of samples. As W increases, the set of ranks given to each object must become more similar because in the error term of (4.3), $\sum_{i=1}^{k} (r_{ij} - \bar{r}_j)^2$ becomes smaller for all j, and thus there is greater agreement between observers. In order to have the interpretation of this k-sample coefficient be consistent with a two-sample measure of association, one might think some measure which ranges from -1 to $+1$ with -1 designating perfect discordance would be preferable. However, for more than two samples, there is no such thing as perfect disagreement between rankings, and thus concordance and discordance are not symmetrical opposites. Therefore the range 0 to 1 is really more appropriate for a k-sample measure of association.

RELATIONSHIP BETWEEN *W* AND RANK CORRELATION

We shall now show that the statistic W is related to the average of the $\binom{k}{2}$ coefficients of rank correlation which can be calculated for the $\binom{k}{2}$ pairings of sample rankings. The average value is

$$r_{\text{av}} = \frac{\sum\sum_{1 \leqslant i<m \leqslant k} r_{i,m}}{\binom{k}{2}} = \sum_{i=1}^{k} \sum_{\substack{m=1 \\ i \neq m}}^{k} \frac{r_{i,m}}{k(k-1)} \tag{4.5}$$

where

$$r_{i,m} = \frac{12}{n(n^2-1)} \sum_{j=1}^{n} \left(r_{ij} - \frac{n+1}{2} \right) \left(r_{mj} - \frac{n+1}{2} \right) \qquad \text{for all } i \neq m$$

Denoting the average rank $(n+1)/2$ by μ, we have

$$\begin{aligned} r_{\text{av}} &= 12 \sum_{j=1}^{n} \sum_{i=1}^{k} \sum_{\substack{m=1 \\ i \neq m}}^{k} \frac{(r_{ij}-\mu)(r_{mj}-\mu)}{kn(k-1)(n^2-1)} \\ &= 12 \sum_{j=1}^{n} \frac{\left[\sum_{i=1}^{k}(r_{ij}-\mu)\right]^2 - \sum_{i=1}^{k}(r_{ij}-\mu)^2}{kn(k-1)(n^2-1)} \\ &= \frac{\sum_{j=1}^{n}(r_j - k\mu)^2 - s_t}{(k-1)s_t} = \frac{s - s_t}{(k-1)s_t} = \frac{kw-1}{k-1} \end{aligned} \tag{4.6}$$

or

$$W = r_{\text{av}} + \frac{1 - r_{\text{av}}}{k} = \frac{r_{\text{av}}(k-1)+1}{k} \tag{4.7}$$

From this relation, we see that $W = 1$ when $r_{\text{av}} = 1$, which can occur only when $r_{i,m}$ equals 1 for all sets (i, m) of two samples, since always $r_{i,m} \leqslant 1$. It is impossible to have $r_{\text{av}} = -1$, since $r_{i,m} = -1$ cannot occur for all sets (i, m) simultaneously. Since we have already shown that the minimum value of W is zero, it follows from (4.7) that the smallest possible value of r_{av} is $-1/(k-1)$.

TESTS OF SIGNIFICANCE BASED ON *W*

Suppose we consider each column in our $k \times n$ table to be the ranks of observations from a k-variate population. With n columns, we can say that $(R_{1j}, R_{2j}, \ldots, R_{kj})$, $j = 1, 2, \ldots, n$, constitute ranks of a random sample of size n from a k-variate population. We wish to test the null hypothesis that the variates are independent. The coefficient of concordance W is an overall measure of the association between the ranks of the k variates or the k sets of rankings of n objects, which in turn estimates some measure of the relationship between the k variates in the population. If the variates are independent, there is no association and W is zero, and for complete dependence there is perfect agreement and W equals 1. Therefore the statistic W may be used to test the null hypothesis that the variates are independent. The appropriate rejection region is large values of W.

In the null case, the ranks assigned to the n observations are completely random for each of the k variates, and the $(n!)^k$ assignments are all equally likely. The random sampling distribution of S (or W) then is exactly the same as in Section 12.2. Table N of the Appendix can therefore be used, and for k large the distribution of

$$Q = \frac{12S}{kn(n+1)} = k(n-1)W$$

may be approximated by the chi-square distribution with $n-1$ degrees of freedom.

Other approximations are also occasionally employed for tests of significance. Although the mean and variance of W are easily found using the moments already obtained for S in (2.7), it will be more instructive to determine the null moments of W directly by using its relationship with R_{av} given in (4.7). From (11.3.13), the mean and variance of $R_{i,m}$, the rank-correlation coefficient of any pairing of independent sets of ranks, are

$$E(R_{i,m}) = 0 \qquad \text{var}(R_{i,m}) = \frac{1}{n-1} \qquad \text{for all } 1 \leqslant i < m \leqslant k$$

For any two independent sets of pairings of independent ranks, say (i,m) and (p,j) where $1 \leqslant i < m \leqslant k,\ 1 \leqslant p < j \leqslant k$, the covariance is

$$\text{cov}(R_{i,m}, R_{p,j}) = 0 \qquad \text{unless } i = p \text{ and } m = j$$

Therefore, from the definition of R_{av} in (4.5), we have

$$E(R_{\text{av}}) = 0$$

$$\binom{k}{2}^2 \text{var}(R_{\text{av}}) = \sum_{1 \leqslant i < m \leqslant k} \text{var}(R_{i,m}) + \sum \sum_{\substack{1 \leqslant i < m \leqslant k \\ i \neq p \text{ or } m \neq j}} \sum_{1 \leqslant p < j \leqslant k} \sum \text{cov}(R_{i,m}, R_{p,j}) = \binom{k}{2} \Big/ (n-1)$$

and

$$\text{var}(R_{\text{av}}) = \frac{2}{k(k-1)(n-1)}$$

Now using (4.7),

$$E(W) = \frac{1}{k} \qquad \text{var}(W) = \frac{2(k-1)}{k^3(n-1)} \tag{4.8}$$

The reader can verify that these are exactly equal to the mean and the variance of the beta distribution with parameters

$$\alpha = \frac{k(n-1)-2}{2k} \qquad \text{and} \qquad \beta = \frac{(k-1)[k(n-1)-2]}{2k}$$

An investigation of the higher moments of W shows that they are approximately equal to the corresponding higher moments of the beta distribution unless $k(n-1)$ is small. Thus an approximation to the distribution of W is the beta distribution, for which tables are available. However, if any random variable, say X, has the beta distribution with parameters α and β, the transformation $Y = \beta X/[\alpha(1-X)]$ produces a random variable with Snedecor's F distribution with parameters $v_1 = 2\alpha$ and $v_2 = 2\beta$, and the transformed variable $Z = (\ln Y)/2$ has Fisher's z distribution with the same parameters. Applying these transformations here, we find the approximate distributions

1. $(k-1)W/(1-W)$ is Snedecor's F with $\nu_1 = n - 1 - 2/k$ and $\nu_2 = (k-1)v_1$
2. $\ln[(k-1)W/(1-W)]/2$ is Fisher's z with $\nu_1 = n-1-2/k$ and $\nu_2 = (k-1)v_1$

in addition to our previous approximation

3. $k(n-1)W$ is chi-square with $n-1$ degrees of freedom.

Approximation 1 is not surprising, as we found in (2.10) that the random variable

$$\frac{(k-1)S}{ks_t - S} = \frac{(k-1)W}{1-W}$$

was the ratio of mean squares analogous to the analysis-of-variance test statistic with $n-1$ and $(n-1)(k-1)$ degrees of freedom.

ESTIMATION OF THE TRUE PREFERENTIAL ORDER OF OBJECTS

Assume that the coefficient of concordance is computed for some k sets of rankings of n objects and the null hypothesis of no agreement is rejected. The magnitude of this relative measure of agreement implies that not all these particular observers ranked the objects strictly randomly and independently. This might be interpreted to mean that there is some agreement among these observers and that perhaps some unique ordering of these objects exists in their estimation. Suppose we call this the true preferential ordering. If there were perfect agreement, we would know which object is least preferred, which is next, etc., by the agreed-upon ranks. Object number j would have the position m in the true preferential ordering if the sum of ranks given object j is km. In our $k \times n$ table of ranks, the ordering corresponds to the ranks of the column sums. In a case of less-than-perfect agreement, then, the true preferential ordering might be estimated by assigning ranks to the objects in accordance with the magnitudes of the column sums.

This estimate is best in the sense that if the coefficient of rank correlation is calculated between this estimated ranking and each of the k observed rankings, the average of these k correlation coefficients is a maximum. To show this, we let $r_{e1}, r_{e2}, \ldots, r_{en}$ be any estimate of the true preferential ordering, where r_{ej} is the estimated rank of object number j. If $R_{e,i}$ denotes the rank-correlation coefficient between this estimated ranking and the ranking assigned by the ith observer, the average rank correlation is

$$\sum_{i=1}^{k} \frac{r_{e,i}}{k} = 12 \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{(r_{ej} - \mu)(r_{ij} - \mu)}{kn(n^2 - 1)} = 12 \sum_{j=1}^{n} \frac{(r_{ej} - \mu)(r_j - k\mu)}{kn(n^2 - 1)}$$

$$= \frac{12 \sum_{j=1}^{n} r_{ej} r_j}{kn(n^2 - 1)} - \frac{3(n+1)}{(n-1)}$$

where $\mu = (n+1)/2$ and r_j is the jth column sum as before. This average then is a maximum when $\sum_{j=1}^{n} r_{ej} r_j$ is a maximum, i.e., when the r_{ej} are in the same relative order of magnitude as the r_j.

This estimate is also best in a least-squares sense. If r_{ej} is any estimated rank of object j and the estimate is the true preferential rank, the jth column sum would equal kr_{ej}. A measure of the error in this estimate then is the sum of squares of deviations

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$\begin{aligned}\sum_{j=1}^{n}(r_j - kr_{ej})^2 &= \sum_{j=1}^{n} r_j^2 + k^2\sum_{j=1}^{n} r_{ej}^2 - 2k\sum_{j=1}^{n} r_j r_{ej} \\ &= \sum_{j=1}^{n} r_j^2 + k^2\sum_{i=1}^{n} i^2 - 2k\sum_{j=1}^{n} r_j r_{ej} \\ &= c - 2k\sum_{j=1}^{n} r_j r_{ej}\end{aligned}$$

where c is a constant. The error is thus minimized when $\sum_{j=1}^{n} r_j r_{ej}$ is a maximum, and the r_{ej} should be chosen as before.

TIED OBSERVATIONS

Up to now we have assumed that each row of our $k \times n$ table is a permutation of the first n integers. If an observer cannot express any preference between two or more objects, or if the objects are actually indistinguishable, we may wish to allow the observer to assign equal ranks. If these numbers are the midranks of the positions each set of tied objects would occupy if a preference could be expressed, the average rank of any object and the average column sum are not changed. However, the sum of squares of deviations of any set of n ranks is reduced if there are ties. As we found in (11.3.18), for any $i = 1, 2, \ldots, k$, the corrected value is

$$\sum_{j=1}^{n}\left(r_{ij} - \frac{n+1}{2}\right)^2 = \frac{n(n^2-1) - \sum t(t^2-1)}{12}$$

The maximum value of s/k, as in (2.9), is then reduced to

$$s_t = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(r_{ij} - \frac{n+1}{2}\right)^2 = \frac{kn(n^2-1) - \sum\sum t(t^2-1)}{12}$$

where the double sum is extended over all sets of t tied ranks in each of the k rows. The relative measure of agreement in the presence of ties then is $W = S/ks_t$. The significance of the corrected coefficient W can be tested using any of the previously mentioned approximations for W.

APPLICATIONS

The coefficient of concordance is a descriptive measure of the agreement between k sets of rankings of n objects and is defined in (4.4)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

where S is easily calculated from (2.11). To test the null hypothesis of no association or no agreement between rankings against the alternative of agreement or positive dependence, Table N of the Appendix can be used to find a right-tail critical value or P value for S in small samples. For large samples, the test statistic Q in (2.8) or equivalently $k(n-1)W$ can be used with Table B and $n-1$ degrees of freedom.

Example 4.1 Eight graduate students are each given examinations in quantitative reasoning, vocabulary, and reading comprehension. Their scores are listed below. It is frequently claimed that persons who excel in quantitative reasoning are not as capable with verbal, and vice versa, and yet a truly intelligent person must possess all of these abilities. Test these data to see if there is an association between scores. Does there seem to be an indirect relationship between quantitative and verbal abilities?

	Student							
Test	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Quantitative	90	60	45	48	58	72	25	85
Vocabulary	62	81	92	76	70	75	95	72
Reading	60	91	85	81	90	76	93	80

Solution The first step is to rank the students from 1 (best) to 8 according to their scores on each of the three skills. This will give us $k=3$ sets of rankings of $n=8$ objects. Then we compute the rank sums as shown below.

	Student							
Test	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Quantitative	1	4	7	6	5	3	8	2
Vocabulary	8	3	2	4	7	5	1	6
Reading	8	2	4	5	3	7	1	6
Total	17	9	13	15	15	15	10	14

For these data $\sum R^2 = 1510$ and S from (2.11) is $S = 52$, and $W = 0.138$ from (4.4) is a descriptive measure of the agreement between rankings. To test the null hypothesis, we need Q from (2.8),

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

which is $Q = 2.89$ with 7 degrees of freedom. The P value from Table B is $0.50 < P < 0.90$, so there appears to be no agreement between the ranks. We might note that the greatest source of disagreement is the Quantitative scores in comparison with the other two, as suggested by the question about an indirect relationship between quantitative and verbal abilities. One way to answer this question statistically is to obtain a verbal score for each student as the sum of the vocabulary and reading scores and compare this ranking with the quantitative ranking using say the rank correlation coefficient. We do this now.

	Student							
Test	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Verbal score	122	171	177	157	160	151	188	152
Verbal rank	8	3	2	5	4	7	1	6
Quantitative rank	1	4	7	6	5	3	8	2

We have $\sum R^2 = 158$ and $R = -0.881$ with $P = 0.004$ from Table M. There is a strong negative dependence between verbal and quantitative scores.

The SPSSX and STATXACT calculations of the Kendall coefficient of concordance are shown below. Note that the statistic value agrees exactly with hand calculations. The packages give the same asymptotic P value, 0.8951, which leads to the same decision as ours. STATXACT provides the exact P value, 0.9267, and the decision is the same.

```
***************************************
   SPSSX SOLUTION TO EXAMPLE 4.1
***************************************

Kendall's W Test

Ranks
             Mean Rank
   VAR00001       3.33
   VAR00002       6.00
   VAR00003       4.67
   VAR00004       4.00
   VAR00005       4.00
   VAR00006       4.00
   VAR00007       5.67
   VAR00008       4.33

Test Statistics
                N        3
      Kendall's W     .138
       Chi-Square    2.889
               df        7
      Asymp. Sig.     .895
Kendall's Coefficient of Concordance
```


MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

```
****************************************
   STATXACT SOLUTION TO EXAMPLE 4.1
****************************************

KENDALL'S CONCORDANCE TEST

Statistic based on the observed by 3 Two-way layout(x):
  W(x) : Kendall Coefficient of Concordance = 0.1376

Asymptotic P-value:(based on Chi-Square distribution with 7df)
  Pr{W(X).GE. 0.1376}=  0.8951

Exact p-value and point probability :
  Pr{W(X).GE. 0.1376}=   0.9267
  Pr{W(X).EQ. 0.1376}=   0.0072
```

12.5 THE COEFFICIENT OF CONCORDANCE FOR *k* SETS OF INCOMPLETE RANKINGS

As an extension of the sampling situation of Section 12.4, suppose that we have n objects to be ranked and a fixed number of observers to rank them but each observer ranks only some subset of the n objects. This situation could arise for reasons of economy or practicality. In the case of human observers particularly, the ability to rank objects effectively and reliably may be a function of the number of comparative judgments to be made. For example, after 10 different brands of bourbon have been tasted, the discriminatory powers of the observers may legitimately be questioned.

We shall assume that the experimental design in this situation is such that the rankings are incomplete in the same symmetrical way as in the balanced incomplete-blocks design which is used effectively in agricultural field experiments. In terms of our situation, this means that:

1. Each observer will rank the same number m of objects for some $m < n$.
2. Every object will be ranked exactly the same total number k of times.
3. Each pair of objects will be presented together to some observer a total of exactly λ times, $\lambda \geqslant 1$, a constant for all pairs.

These specifications then ensure that all comparisons are made with the same frequency.

In order to visualize the design, imagine a two-way layout of p rows and n columns, where the entry δ_{ij} in *(i,j)* cell equals 1 if object j is presented to observer i and 0 otherwise. The design specifications then can be written symbolically as

1. $\sum_{j=1}^{n} \delta_{ij} = m \qquad$ for $i = 1, 2, \ldots, p$
2. $\sum_{i=1}^{p} \delta_{ij} = k \qquad$ for $j = 1, 2, \ldots, n$
3. $\sum_{i=1}^{p} \delta_{ij}\delta_{ir} = \lambda \qquad$ for all $r \neq j = 1, 2, \ldots, n$

Summing on the other subscript in specifications 1 and 2, we obtain

$$\sum_{i=1}^{p} \sum_{j=1}^{n} \delta_{ij} = mp = kn$$

which implies that the number of observers is fixed by the design to be $p = kn/m$. Now using specification 3, we have

$$\sum_{i=1}^{p} \left(\sum_{j=1}^{n} \delta_{ij} \right)^2 = \sum_{i=1}^{p} \left(\sum_{j=1}^{n} \delta_{ij}^2 + \sum_{\substack{j=1 \\ j \neq r}}^{n} \sum_{r=1}^{n} \delta_{ij}\,\delta_{ir} \right) = mp + \lambda n(n-1)$$

and from specification 1, this same sum equals pm^2. This requires the relation

$$\lambda = \frac{pm(m-1)}{n(n-1)} = \frac{k(m-1)}{n-1}$$

Since p and λ must both be positive integers, m must be a factor of kn and $n-1$ must be a factor of $k(m-1)$. Designs of this type are called *Youden squares* or *incomplete Latin squares*. Such plans have been tabulated (for example, in Cochran and Cox, 1957, pp. 520–544). An example of this design for $n = 7, \lambda = 1, m = k = 3$, where the objects are designated by *A, B, C, D, E, F*, and *G* is:

Observer	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Objects presented for ranking	*A*	*B*	*C*	*D*	*E*	*F*	*G*
	B	*C*	*D*	*E*	*F*	*G*	*A*
	D	*E*	*F*	*G*	*A*	*B*	*C*

We are interested in determining a single measure of the overall concordance or agreement between the kn/m observers in their relative comparisons of the objects. For simplification, suppose there is some natural ordering of all n objects and the objects labeled accordingly. In other words, object number r would receive rank r by all observers if each observer was presented with all n objects and the observers agreed perfectly in their evaluation of the objects. For perfect agreement in a balanced incomplete ranking then, where each observer assigns ranks $1, 2, \ldots, m$ to the subset presented to him, object 1 will receive rank 1 whenever it is presented; object 2 will receive rank 2 whenever it is presented along with object 1, and rank 1 otherwise; object 3 will receive rank 3 when presented along with both objects 1 and 2, rank 2 when with either objects 1 or 2 but not both, and rank 1 otherwise, etc. In general, then, the rank of object j when presented to observer i is one more than the number of objects presented to that observer from the subset of objects $\{1, 2, \ldots, j-1\}$, for all $2 \leqslant j \leqslant n$. Symbolically, using the δ notation of before, the rank of object j when presented to observer i is 1 for $j = 1$ and

$$1 + \sum_{r=1}^{j-1} \delta_{ir} \qquad \text{for all } 2 \leqslant j \leqslant n$$

The sum of the ranks assigned to object j by all p observers in the case of perfect agreement then is

$$\sum_{i=1}^{p} \left(1 + \sum_{r=1}^{j-1} \delta_{ir}\right) \delta_{ij} = \sum_{i=1}^{p} \delta_{ij} + \sum_{r=1}^{j-1} \sum_{i=1}^{p} \delta_{ir}\, \delta_{ij} \bar{k} + \lambda(j-1)$$

$$\text{for } j = 1, 2, \ldots, n$$

as a result of the design specifications 2 and 3.

Since each object is ranked a fixed number, k, of times, the observed data for an experiment of this type can easily be presented in a two-way layout of k rows and n columns, where the jth column contains the collection of ranks assigned to object j by those observers to whom object j was presented. The rows no longer have any significance, but the column sums can be used to measure concordance. The sum of all ranks in the table is $[m(m+1)/2][kn/m] = kn(m+1)/2$, and thus the average column sum is $k(m+1)/2$. In the case of perfect concordance, the column sums are some permutation of the numbers

$$k,\ k+\lambda, k+2\lambda, \ldots, k+(n-1)\lambda$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

and the sums of squares of deviations of column sums around their mean is

$$\sum_{j=0}^{n-1}\left[(k+j\lambda)-\frac{k(m+1)}{2}\right]^2=\frac{\lambda^2 n(n^2-1)}{12}$$

Let R_j denote the actual sum of ranks in the jth column. A relative measure of concordance between observers may be defined here as

$$W=\frac{12\sum_{j=1}^{n}[R_j-k(m+1)/2]^2}{\lambda^2 n(n^2-1)} \tag{5.1}$$

If $m=n$ and $\lambda=k$ so that each observer ranks all n objects, (5.1) is equivalent to (4.4), as it should be.

This coefficient of concordance also varies between 0 and 1 with larger values reflecting greater agreement between observers. If there is no agreement, the column sums would all tend to be equal to the average column sum and W would be zero.

TESTS OF SIGNIFICANCE BASED ON *W*

For testing the null hypothesis that the ranks are allotted randomly by each observer to the subset of objects presented to him so that there is no concordance, the appropriate rejection region is large values of W. This test is frequently called the *Durbin (1951) test*.

The exact sampling distribution of W could be determined only by an extensive enumeration process. Exact tables for 15 different designs are given in van der Laan and Prakken (1972). For k large an approximation to the null distribution may be employed for tests of significance. We shall first determine the exact null mean and variance of W using an approach analogous to the steps leading to (2.7). Let R_{ij}, $i=1,2,\ldots,k$, denote the collection of ranks allotted to object number j by the k observers to whom it was presented. From (11.3.2), (11.3.3), and (11.3.10), in the null case then for all i, j, and $q \neq j$

$$E(R_{ij})=\frac{m+1}{2} \qquad \operatorname{var}(R_{ij})=\frac{m^2-1}{12} \qquad \operatorname{cov}(R_{ij},R_{iq})=-\frac{m+1}{12}$$

and R_{ij} and R_{hj} are independent for all j where $i \neq h$. Denoting $(m+1)/2$ by μ, the numerator of W in (5.1) may be written as

$$
\begin{aligned}
&12\sum_{j=1}^{n}\left[\sum_{i=1}^{k}R_{ij}-k\mu\right]^2\\
&\quad=12\sum_{j=1}^{n}\left[\sum_{i=1}^{k}(R_{ij}-\mu)\right]^2\\
&\quad=12\sum_{j=1}^{n}\sum_{i=1}^{k}(R_{ij}-\mu)^2+24\sum_{j=1}^{n}\sum\sum_{1\leqslant i<h\leqslant k}(R_{ij}-\mu)(R_{hj}-\mu)\\
&\quad=pm(m^2-1)+24U=\lambda^2n(n^2-1)W \qquad (5.2)
\end{aligned}
$$

Since $\operatorname{cov}(R_{ij},R_{hj})=0$ for all $i<h$, $E(U)=0$. Squaring the sum represented by U, we have

$$
\begin{aligned}
U^2=&\sum_{j=1}^{n}\sum\sum_{1\leqslant i<h\leqslant k}(R_{ij}-\mu)^2(R_{hj}-\mu)^2+2\sum\sum_{1\leqslant j<q\leqslant n}\sum\sum_{1\leqslant i<h\leqslant k}\\
&\times\sum\sum_{1\leqslant r<s\leqslant k}(R_{ij}-\mu)(R_{hj}-\mu)(R_{rq}-\mu)(R_{sq}-\mu)
\end{aligned}
$$

and

$$
\begin{aligned}
E(U^2)=&\sum_{j=1}^{n}\sum\sum_{1\leqslant i<h\leqslant k}\operatorname{var}(R_{ij})\operatorname{var}(R_{hj})\\
&+2\sum\sum_{1\leqslant j<q\leqslant n}\binom{\lambda}{2}\operatorname{cov}(R_{ij},R_{iq})\operatorname{cov}(R_{hj},R_{hq})
\end{aligned}
$$

since objects j and q are presented together to both observers i and h a total of $\binom{\lambda}{2}$ times in the experiment. Substituting the respective variances and covariances, we obtain

$$
\begin{aligned}
\operatorname{var}(U)=E(U^2)&=\frac{n\binom{k}{2}(m^2-1)^2+2\binom{n}{2}\binom{\lambda}{2}(m+1)^2}{144}\\
&=nk(m+1)^2(m-1)\frac{(m-1)(k-1)+(\lambda-1)}{288}
\end{aligned}
$$

From (5.2), the moments of W are

$$
E(W)=\frac{m+1}{\lambda(n+1)}
$$

$$
\operatorname{var}(W)=2(m+1)^2\frac{(m-1)(k-1)+(\lambda-1)}{nk\lambda^2(m-1)(n+1)^2}
$$

As in the case of complete rankings, a linear function of W has moments approximately equal to the corresponding moments of the chi-square distribution with $n-1$ degrees of freedom if k is large. This function is

$$Q = \frac{\lambda(n^2-1)W}{m+1}$$

and its exact mean and variance are

$$E(Q) = n-1$$

$$\operatorname{var}(Q) = 2(n-1)\left[1 - \frac{m(n-1)}{nk(m-1)}\right] \approx 2(n-1)\left(1 - \frac{1}{k}\right)$$

The rejection region for large k and significant level α then is

$$Q \in R \qquad \text{for } Q \geqslant \chi^2_{n-1,\alpha}$$

TIED OBSERVATIONS

Unlike the case of complete rankings, no simple correction factor can be introduced to account for the reduction in total sum of squares of deviations of column totals around their mean when the midrank method is used to handle ties. If there are only a few ties, the null distribution of W should not be seriously altered, and thus the statistic can be computed as usual with midranks assigned. Alternatively, any of the other methods of handling ties discussed in Section 5.6 (except omission of tied observations) may be adopted.

APPLICATIONS

This analysis-of-variance test based on ranks for balanced incomplete rankings is usually called the Durbin test. The test statistic here, where λ is the number of times each pair of treatments is ranked and m is the number of treatments in each block, is most easily computed as

$$Q = \frac{12\sum_{j=1}^{n} R_j^2}{\lambda n(m+1)} - \frac{3k^2(m+1)}{\lambda} \tag{5.3}$$

which is asymptotically chi-square distributed with $n-1$ degrees of freedom. The null hypothesis of equal treatment effects is rejected for Q large.

Kendall's coefficient of concordance descriptive measure for k incomplete sets of n rankings, where m is the number of objects presented

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

for ranking and λ is the number of times each pair of objects is ranked together, is given in (5.1), which is equivalent to

$$W = \frac{12\sum_{j=1}^{n} R_j^2 - 3k^2 n(m+1)^2}{\lambda^2 n(n^2-1)} \tag{5.4}$$

and $Q = \lambda(n^2-1)W/(m+1)$ is the chi-square test statistic with $n-1$ degrees of freedom for the null hypothesis of no agreement between rankings.

If the null hypothesis of equal treatment effects is rejected, we can use a multiple comparisons procedure to determine which pairs of treatments have significantly different effects. Treatments i and j are declared to be significantly different if

$$|R_i - R_j| \geqslant z^* \sqrt{\frac{km(m^2-1)}{6(n-1)}} \tag{5.5}$$

where z^* is the negative of the $[\alpha/n(n-1)]$th quantile of the standard normal distribution.

Example 5.1 A taste-test experiment to compare seven different kinds of wine is to be designed such that no taster will be asked to rank more than three different kinds, so we have $n = 7$ and $m = 3$. If each pair of wines is to be compared only once so that $\lambda = 1$, the required number of tasters is $p = \lambda n(n-1)/m(m-1) = 7$. A balanced design was used and the rankings given are shown below. Calculate Kendall's coefficient of concordance as a measure of agreement between rankings and test the null hypothesis of no agreement.

	Wine						
Taster	*A*	*B*	*C*	*D*	*E*	*F*	*G*
1	1	2		3			
2		1	3		2		
3			3	2		1	
4				2	3		1
5	1				3	2	
6		2				1	3
7	1		3				2
	—	—	—	—	—	—	—
Total	3	5	9	7	8	4	6

Solution Each wine is ranked three times so that $k = 3$. We calculate $\sum R_j^2 = 280$ and substitute into (5.4) to get $W = 1$, which describes

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

perfect agreement. The test statistic from (5.3) is $Q = 12$ with 6 degrees of freedom. The P value from Table B of the Appendix is $0.05 < P < 0.10$ for the test of no agreement between rankings. At the time of this writing, neither STATXACT nor SAS has an option for the Durbin test.

12.6 KENDALL'S TAU COEFFICIENT FOR PARTIAL CORRELATION

Coefficients of partial correlation are useful measures for studying relationships between two random variables since they are ordinary correlations between two variables with the effects of some other variables eliminated because these latter variables are held constant. In other words, the coefficients measure association in the conditional probability distribution of two variables given one or more other variables. A nice property of Kendall's tau coefficient of Section 11.2 is that it can be easily extended to the theory of partial correlation.

Assume we are given m independent observations of triplets $(X_i, Y_i, Z_i), i = 1, 2, \ldots, m$, from a trivariate population where the marginal distributions of each variable are continuous. We wish to determine a sample measure of the association between X and Y when Z is held constant. Define the indicator variables

$$U_{ij} = \text{sgn}(X_j - X_i) \qquad V_{ij} = \text{sgn}(Y_j - Y_i) \qquad W_{ij} = \text{sgn}(Z_j - Z_i)$$

and for all $1 \leqslant i < j \leqslant m$, let $n(u,v,w)$ denote the number of values of (i,j) such that $u_{ij} = u, v_{ij} = v, w_{ij} = w$. Now we further define the count variables

$$X_{11} = n(1, 1, 1)$$
$$X_{22} = n(-1, -1, 1)$$
$$X_{12} = n(-1, 1, 1)$$
$$X_{21} = n(1, -1, 1)$$

Then X_{11} is the number of sets of (i,j) pairs, $1 \leqslant i < j \leqslant m$, of each variable such that X and Y are both concordant with Z, X_{22} is the number where X and Y are both discordant with Z, X_{12} is the number such that X is discordant with Z and Y is concordant with Z, and X_{21} is the number where X is concordant with Z and Y is discordant with Z. We present these counts in a 2×2 table as shown in Table 6.1. This table sets out the agreements of rankings X with Z, and rankings Y with Z, and the same for the disagreements. Now we define the partial rank correlation coefficient between X and Y when Z is held constant as

$$T_{XY.Z} = \frac{X_{11}X_{22} - X_{12}X_{21}}{(X_{.1}X_{.2}X_{1.}X_{2.})^{1/2}} \tag{6.1}$$

The value of this coefficient ranges between -1 and $+1$. At either of these two extremes, we have

$$(X_{11}+X_{21})(X_{12}+X_{22})(X_{11}+X_{12})(X_{21}+X_{22}) - (X_{11}X_{22} - X_{12}X_{21})^2 = 0$$

a sum of products of three or more nonnegative numbers whose exponents total four equal to zero. This occurs only if at least two of the numbers are zero. If $X_{ij} = X_{hk} = 0$ for $i{=}h$ or $j = k$, either X or Y is in perfect concordance or discordance with Z. The nontrivial cases then are where both diagonal entries are zero. If $X_{12} = X_{21} = 0$, the X and Y sample values are always either both concordant or both discordant with Z and $T_{XY.Z} = 1$. If $X_{11} = X_{22} = 0$, they are never both in the same relation and $T_{XY.Z} = -1$. Maghsoodloo (1975, 1981) and Moran (1951) give tables of the sampling distribution of the partial tau coefficient. $T_{XY.Z}$ provides a useful relative measure of the degree to which X and Y are concordant when their relation with Z is eliminated.

It is interesting to look at the partial tau coefficient in a different algebraic form. Using the X_{ij} notation above, the Kendall tau coefficients for the three different paired samples would be

$$\binom{m}{2}T_{XY} = (X_{11}+X_{22}) - (X_{12}+X_{21})$$

$$\binom{m}{2}T_{XZ} = (X_{11}+X_{21}) - (X_{22}+X_{12})$$

$$\binom{m}{2}T_{YZ} = (X_{11}+X_{12}) - (X_{22}+X_{21})$$

Since $\binom{m}{2} = X_{11}+X_{12}+X_{12}+X_{22} = n$, we have

Table 6.1 Presentation of data

	Ranking X		
Ranking Y	*Pairs Concordant with Z*	*Pairs Discordant with Z*	*Total*
Pairs Concordant with Z	X_{11}	X_{12}	$X_{1.}$
Pairs Discordant with Z	X_{21}	X_{22}	$X_{2.}$
Total	$X_{.1}$	$X_{.2}$	$X_{..} = N$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$1 - T_{XZ}^2 = \frac{4(X_{11} + X_{21})(X_{12} + X_{22})}{n^2} = \frac{4X_{.1}X_{.2}}{n^2}$$
$$1 - T_{YZ}^2 = \frac{4(X_{11} + X_{12})(X_{22} + X_{21})}{n^2} = \frac{4X_{1.}X_{2.}}{n^2}$$

and

$$n^2 T_{XY} = [(X_{11} + X_{22}) - (X_{12} + X_{21})][(X_{11} + X_{22}) + (X_{12} + X_{21})]$$
$$n^2(T_{XY} - T_{XZ}T_{YZ}) = 4(X_{11}X_{22} - X_{12}X_{21})$$

Therefore (6.1) can be written as

$$T_{XY.Z} = \frac{T_{XY} - T_{XZ}T_{YZ}}{[(1 - T_{XZ}^2)(1 - T_{YZ}^2)]^{1/2}} \tag{6.2}$$

Some other approaches to defining a measure of partial correlation have appeared in the journal literature. One of the more useful measures is the index of matched correlation proposed by Quade (1967).

The partial tau defined here has a particularly appealing property in that it can be generalized to the case of more than three variables. Note that the form in (6.2), with each T replaced by its corresponding R, is identical to the expression for a Pearson product-moment partial correlation coefficient. This is because both are special cases of a generalized partial correlation coefficient which is discussed in Somers (1959). With his generalized form, extensions of the partial tau coefficient to higher orders are possible.

APPLICATIONS

The null hypothesis to be tested using $T_{XY.Z}$ in (6.2) is that X and Y are independent when the effect of Z is removed. The appropriate rejection regions are large values of $T_{XY.Z}$ for the alternative of positive dependence and small values for the alternative of negative dependence. The null distribution of $T_{XY.Z}$ is given in Table P of the Appendix as a function of the number of rankings m.

Example 6.1 Maghsoodloo (1975) used an example with $m = 7$ sets of rankings on three variables. The data are given in Table 6.2, arranged so that the ranking of the Z variable follows the natural order.

Table 6.2 Data for Example 6.1

	Subject						
Variable	*B*	*D*	*C*	*A*	*E*	*G*	*F*
Z	1	2	3	4	5	6	7
X	6	7	5	3	4	1	2
Y	7	6	5	3	4	2	1

Compute the partial correlation between X and Y given Z and test for positive dependence. Compare the result with that for X and Y when the effect of Z is not removed.

Solution We compute Kendall's tau coefficient between each set of pairs $(X, Y), (X, Z)$, and (Y, Z) in the usual way. For X and Z, the number of concordant pairs is $C = 3$ and the number of discordant pairs is $Q = 18$, giving $T_{XZ} = -0.7143$. For Y and Z, $C = 1$ and $Q = 20$ with $T_{YZ} = -0.9048$. For X and Y, $C = 19$ and $Q = 2$ with $T_{XY} = 0.8095$ and $P = 0.005$ from Table L with $m = 7$, so there is a positive association between X and Y.

Now we compute the partial tau from (6.2) as

$$T_{XY.Z} = \frac{0.8095 - (-0.7143)(-0.9048)}{\sqrt{[1 - (-0.7143)^2][1 - (-0.9048)^2]}} = 0.548$$

From Table P the one-tailed P value is between 0.025 and 0.05. The positive association previously observed between X and Y is much weaker when the effect of Z is removed.

12.7 SUMMARY

In this chapter we have covered a number of different descriptive and inferential procedures involving measures of association in multiple classifications. First, in Section 12.2 we presented Friedman's test for equal treatment effects in the two-way analysis-of-variance for the completely randomized design with k blocks and n treatments. This design is frequently called the *repeated measures design* in the behavioral and social sciences literature. If the null hypothesis of equal treatment effects is rejected, we have a multiple comparisons procedure to determine which pairs of treatments differ and in which direction, with one overall level of significance. If the alternative states an *a priori* order for the treatment effects, we can use Page's test for ordered alternatives, covered in Section 12.3. If the randomized block design is incomplete in a balanced way, so that all treatments are not observed in each block but the presentation is balanced, we can use the Durbin test covered in Section 12.5. A multiple comparisons test is also available to compare the treatments in this design.

The topic covered in Section 12.4 is measures of association for k sets of rankings of n objects. The descriptive measure is Kendall's coefficient of concordance, which ranges between zero and one, with increasing values reflecting increasing agreement among the k rankings. When there is a significant agreement among the rankings, we can estimate the overall "agreed upon" preference in accordance with the sample rank totals for the n objects. This is the least-squares estimate. We found a linear relationship between this coefficient of concordance and the average of the Spearman rank correlation coefficients that could have been calculated for all of the $k(k-1)/2$ pairs of rankings. This situation is extended to the case of k sets of incomplete rankings of n objects in Section 12.5. Then we covered in Section 12.6 the topic of partial correlation for three rankings of n objects. Here the Kendall coefficient of partial correlation measures the association between two variables when the effect of a third variable has been removed or "averaged out." This descriptive measure ranges between -1 and $+1$, with increasing absolute values reflecting a greater degree of association or dependence between variables.

PROBLEMS

12.1. Four varieties of soybean are each planted in three blocks. The yields are:

	Variety of soybean			
Block	*A*	*B*	*C*	*D*
1	45	48	43	41
2	49	45	42	39
3	38	39	35	36

Use Friedman's analysis of variance by ranks to test the hypothesis that the four varieties of soybean all have the same effect on yield.

12.2. A beauty contest has eight contestants. The three judges are each asked to rank the contestants in a preferential order of pulchritude. The results are:

	Contestant							
Judge	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
1	2	1	3	5	4	8	7	6
2	1	2	4	5	7	6	8	3
3	3	2	1	4	5	8	7	6

(*a*) Calculate Kendall's coefficient of concordance between rankings.

(*b*) Calculate the coefficient of rank correlation for each of the three pairs of rankings and verify the relation between r_{av} and W given in (4.7).

(*c*) Estimate the true preferential order of pulchritude.

12.3. Derive by enumeration the exact null distribution of W for three sets of rankings of two objects.

12.4. Given the following triplets of rankings of six objects:

X	1	3	5	6	4	2
Y	1	2	6	4	3	5
Z	2	1	5	4	6	3

(*a*) Calculate the Kendall coefficient of partial correlation between X and Y from (6.1) and test for independence.

(*b*) Calculate (6.2) for these same data to verify that it is an equivalent expression.

12.5. Howard, Murphy, and Thomas (1986) (see Problems 5.12 and 8.8) also wanted to determine whether there is a direct relationship between computer anxiety and math anxiety. Even though the two subjects involve somewhat different skills (clear, logical, and serial thinking versus quantitative talent), both kinds of anxiety are frequently present in persons who regard themselves as technologically alienated. The pretest scores are shown in Table 1 for 14 students, with larger scores indicating greater amounts of the trait.

(*a*) Determine the relationship between computer anxiety and math anxiety.

(*b*) Determine the relationship when the effect of technological alienation is removed.

12.6. Webber (1990) reported results of a study to measure optimism and cynicism about the business environment and ethical trends. Subjects, ranging from high

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 1 Data for Problem 12.5

Student	*Math anxiety*	*Computer anxiety*	*Technological alienation*
A	20	22	18
B	21	24	20
C	23	23	19
D	26	28	25
E	32	34	36
F	27	30	28
G	38	38	42
H	34	36	40
I	28	29	28
J	20	21	23
K	29	32	32
L	22	25	24
M	30	31	37
N	25	27	25

school students to executives, were asked to respond to a questionnaire with general statements about ethics. Two questions related to subjects' degree of agreement (5-point scale) with general statements about ethics. Three questions related to how others would behave in specific problematic situations and answers were multiple choice. Three more questions, also multiple choice, related to how subjects themselves would react to the same problematic situations. These answers were used to develop an optimism index, where larger numbers indicate an optimistic feeling about current and future ethical conditions, and a cynicism index that measures how subjects felt others would behave relative to the way they themselves would behave (a cynicism index of 2.0, for example, means subjects judged others twice as likely as themselves to engage in unethical behavior). The author claimed an inverse relationship between optimism and cynicism but also noted a relation to organizational status of respondents as measured by age. Use the data in Table 2 to determine whether the relationship between optimism and cynicism is still present when the effect of age is removed.

Table 2 Data for Problem 12.6

Group	*Mean age*	*Optimism index*	*Cynicism index*
Owners/managers	60 +	55	1.1
Corporate executives	44	59	1.4
Middle managers	34	41	1.4
MBA students	25	30	1.8
Undergraduates	20	23	2.2

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

12.7. Eight students are given examinations on each of verbal reasoning, quantitative reasoning, and logic. The scores range from 0 to 100, with 100 a perfect score. Use the data below to find the Kendall partial tau coefficient between quantitative and logic when the effect of verbal is removed. Find the P value. Compare the result to the P value for quantitative and logic alone and interpret the comparison.

	Student							
Score	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Verbal	90	60	45	48	58	72	25	85
Quantitative	62	81	92	76	70	75	95	72
Logic	60	91	85	81	90	76	93	80

12.8. *Automobile Magazine* (July 1989) published results of a comparison test of 15 brand models of \$20,000 sedans. Each car was given a subjective score out of a possible 60 points (60 = best) on each of 10 characteristics that include factors of appearance, comfort, and performance. The scores of the six best models are shown below. Determine whether the median scores are the same. Which model(s) would you buy, on the basis of this report?

	Model					
Factor	*Nissan Maxima SE*	*Acura Legend*	*Toyota Cressida*	*Mitsubishi Galant GS*	*Peugeot 405Mi16*	*Ford Taurus*
Exterior styling	55	40	38	32	43	41
Interior comfort	50	50	47	46	41	40
Fit and finish	51	54	53	49	39	42
Engine	53	51	53	48	38	58
Transmission	60	54	47	45	44	38
Steering	43	42	45	45	57	45
Handling	45	46	43	49	54	45
Quality of ride	48	48	50	43	44	41
Fun to drive	51	46	40	49	53	51
Value for money	53	51	47	53	42	47
Total	509	482	463	459	455	448

12.9. A manufacturer of ice cream carried out a taste preference test on seven varieties of ice cream, denoted by *A, B, C, D, E, F, G*. The subjects were a random sample of 21 tasters and each taster had to compare only three varieties. Each pair of varieties is presented together three times, to a group of seven of the tasters, with the design shown below used each of the three times.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Taster	Varieties presented		
1	*A*	*B*	*D*
2	*B*	*C*	*E*
3	*C*	*D*	*F*
4	*D*	*E*	*G*
5	*E*	*F*	*A*
6	*F*	*G*	*B*
7	*G*	*A*	*C*

The ranks resulting from the three repetitions are shown below with each rank corresponding to the variety shown above. For example, the rank 3 by taster 12 is for variety *E*. Determine whether there is a positive association between the rankings.

Taster	Ranks			Taster	Ranks			Taster	Ranks		
1	2	1	3	8	2	1	3	15	2	1	3
2	1	2	3	9	2	1	3	16	1	2	3
3	2	3	1	10	1	3	2	17	3	2	1
4	3	2	1	11	3	2	1	18	2	1	3
5	3	1	2	12	3	1	2	19	2	1	3
6	1	3	2	13	2	3	1	20	2	1	3
7	1	3	2	14	1	3	2	21	2	1	3

12.10. Ten graduate students take identical comprehensive examinations in their major field. The grading procedure is that each professor ranks each student's paper in relation to all others taking the examination. Suppose that four professors give the following ranks, where 1 indicates the best paper and 10 the worst.

	Student									
Professor	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	5	3	8	9	2	7	6	1	4	10
2	7	4	6	2	3	9	8	5	1	10
3	3	5	7	6	4	10	8	2	1	9
4	4	5	7	8	3	9	6	1	2	10

(a) Is there evidence of agreement among the four professors?
(b) Give an overall estimate of the relative performance of each student.
(c) Will it be difficult to decide which students should be given a passing grade?

12.11. Show that if $m = n$ and $\lambda = k$ in (5.3) so that the design is complete, then (5.3) is equivalent to $Q = 12S/kn(n+1)$, as it should be.

12.12. A town has 10 different supermarkets. For each market, data are available on the following three variables: $X_1 = $ food sales, $X_2 = $ nonfood sales, and $X_3 = $ size of store in thousands of square feet. Calculate the partial tau coefficient for X_1 and X_2, when the effects of X_3 are eliminated.

Store no.	*Size of store (1000 ft^2)*	*Food sales ($10,000)*	*Nonfood sales ($10,000)*
1	35	305	35
2	22	130	98
3	27	189	83
4	16	175	76
5	28	101	93
6	46	269	77
7	56	421	44
8	12	195	57
9	40	282	31
10	32	203	92

12.13. Suppose in Problem 11.15 that an independent group of female consumers also ranked the products as follows:

Product	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Independent female ranks	8	9	5	6	1	2	7	4	10	3

(*a*) Is there agreement between the three sets of rankings? Give a descriptive measure of agreement and find a P value.

(*b*) Use all the data given to estimate the rank ordering of the products. In what sense is this estimate a good one?

12.14. An experimenter is attempting to evaluate the relative effectiveness of four drugs in reducing the pain and trauma of persons suffering from migraine headaches. He gave seven patients each drug for a month at a time. At the end of each month, each patient gave an estimate of the relative degree of pain suffered from migraines during that month on a scale from 0 to 10, with 10 being the most severe pain. Test the null hypothesis that the drugs are equally effective.

Drug	*1*	*2*	*3*	*4*	*5*	*6*	*7*
A	7	10	7	9	8	8	8
B	7	6	5	8	7	5	7
C	3	7	3	5	4	6	3
D	4	3	2	1	0	1	0

12.15. A study was made of a sample of 100 female students at a large college on the relative popularity of four experimental types of service clubs having essentially the same goals. The types differed only with respect to the difficulty of achieving membership, with type I having no membership requirements, . . . , and type IV having very rigid and formidable membership requirements. The 100 students were assigned randomly into four groups and each student received a letter asking her to come for an interview as a prospective member of the club. At each interview, the goals and membership requirements were outlined and the student was asked to rank on a 10-point scale how eager she was to join the club described (1 = most eager, 10 = least eager). The students in group I were told about type I club, group II about type II, group III about type III, and group IV about type IV, in order to make recording the data easier. The data for the ratings of the 100 students are shown below.

	Group			
Rating	*I*	*II*	*III*	*IV*
1	0	0	0	0
2	0	0	3	1
3	0	2	4	1
4	2	2	5	8
5	3	6	3	10
6	5	5	1	0
7	5	5	4	4
8	7	3	4	1
9	3	2	1	0
10	0	0	0	0
Sum	25	25	25	25

The experimenter comes to you and asks you to:

(*a*) Comment about the design of the experiment, including any criticisms.

(*b*) Help him get some useful information from his study. Do the best you can to help. You may want to use more than one kind of analysis.

13
Asymptotic Relative Efficiency

13.1 INTRODUCTION

In Chapter 1 the concept of Pittman efficiency was defined as a criterion for the comparison of any two test statistics. Many of the nonparametric tests covered in this book can be considered direct analogs of some classical test which is known to be most powerful under certain specific distribution assumptions. The asymptotic efficiencies of the nonparametric tests relative to a “best” test have been simply stated here without discussion. In this chapter we shall investigate the concept of efficiency more thoroughly and prove some theorems which simplify the calculation. The theory will then be illustrated by applying it to various tests covered in earlier chapters in order to derive numerical values of the ARE for some particular distributions. The theory presented here is generally attributed to Pitman; Noether (1955) gives important generalizations of the theory.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Suppose that we have two test statistics, denoted by T and T^*, which can be used for similar types of inferences regarding simple hypotheses. One method of comparing the performance of the two tests was described in Chapter 1 as relative power efficiency. The power efficiency of test T relative to test T^* is defined as the ratio n^*/n, where n^* is the sample size necessary to attain the power γ at significance level α when test T^* is used, and n is the sample size required by test T to attain the same values γ and α.

As a simple numerical example, consider a comparison of the normal-theory test T^* and the ordinary sign test T for the respective hypothesis-testing situations

$$H_0\colon \mu = 0 \qquad \text{versus} \qquad H_1\colon \mu = 1$$

and

$$H_0\colon M = 0 \qquad \text{versus} \qquad H_1\colon M = 1$$

The inference is to be based on a single random sample from a population which is assumed to be normally distributed with known variance equal to 1. Then the hypothesis sets above are identical. Suppose we are interested in the relative sample sizes for a power of 0.90 and a significance level of 0.05. For the most powerful (normal-theory) test based on n^* observations, the null hypothesis is rejected when $\sqrt{n^*}\bar{X} \geqslant 1.64$ for $\alpha = 0.05$. Setting the power γ equal to 0.90, n^* is found as follows:

$$\begin{aligned}\text{Pw}(1) &= \gamma(1)\\ &= P(\sqrt{n^*}\bar{X} \geqslant 1.64 \mid \mu = 1)\\ &= P[\sqrt{n^*}(\bar{X} - 1) \geqslant 1.64 - \sqrt{n^*}] = 0.90\end{aligned}$$

$$\Phi(1.64 - \sqrt{n^*}) = 0.10 \qquad 1.64 - \sqrt{n^*} = -1.28 \qquad n^* = 9$$

The sign test T of Section 5.4 has rejection region $K \geqslant k_\alpha$, where K is the number of positive observations X_i and k_α is chosen so that

$$\sum_{k=k_\alpha}^{n} \binom{n}{k} 0.5^n = \alpha \tag{1.1}$$

The power of the test T then is

$$\sum_{k=k_\alpha}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} = \gamma(1) \tag{1.2}$$

where $\theta = P(X > 0 \mid M = 1) = 1 - \Phi(-1) = 0.8413$, since the mean and median coincide for a normal population. The number n and $k_{0.05}$ will be those values of n and k_α, respectively, which simultaneously satisfy (1.1) and (1.2) when $\alpha = 0.05$ and $\gamma = 0.90$. If θ is rounded off to 0.85, ordinary tables of the binomial distribution can be used instead of actual calculations. Some of the steps relevant to finding the simultaneous solution are shown in Table 1.1.

If we do not wish to resort to the use of randomized decision rules, we can either (1) choose values for n and k_α such that α and γ are both as close as possible to the preselected numbers or (2) choose the smallest value of n such that the smallest value of k_α gives α and $\beta = 1 - \gamma$ no larger than the preselected numbers. We obtain $n = 13$ and $k_{0.05} = 10$ using method 1 and $n = 16$, $k_{0.05} = 12$ with method 2. These methods are undesirable for a number of obvious reasons, but mainly because method 1 may not lead to a unique answer and method 2 may be too conservative with respect to both types of errors. A preferable approach for the purposes of comparison would be to use randomized decision rules. Then we can either make exact $\alpha = 0.05$ or exact $\gamma = 0.90$ but probably not both. When deciding to make exact $\alpha = 0.05$ and $\gamma = 0.90$, the procedure is also illustrated in Table 1.1. Starting with the smallest n and the corresponding smallest k_α for

Table 1.1 Power calculations

				Randomized decision rule for exact $\alpha = 0.05$	
n	k_α	α	$\gamma = 1 - \beta$	*Probability of rejection*	$\gamma(1)$
17	13	0.0245	0.9013		
	12	0.0717	0.9681		
16	12	0.0383	0.9211	1	
15	11	0.1050	0.9766	0.1754	0.9308
	12	0.0176	0.8226	1	
14	11	0.0593	0.9382	0.8010	0.9151
	11	0.0288	0.8535	1	
13	10	0.0899	0.9533	0.3470	0.8881
	10	0.0461	0.8820		
	9	0.1334	0.9650		

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

which simultaneously $\alpha \leqslant 0.05$ and $\gamma \geqslant 0.90$, the randomized decision rule is found by solving for p in the expression

$$\sum_{k=k_\alpha}^{n} \binom{n}{k} 0.5^n + p \binom{n}{k_\alpha - 1} 0.5^n = 0.05$$

Then the power for this exact 0.05 size test is

$$\sum_{k=k_\alpha}^{n} \binom{n}{k} (0.85)^k (0.85)^{n-k} + p \binom{n}{k_\alpha - 1} (0.85)^k (0.15)^{n-k}$$

Do the same set of calculations for the next smaller n, etc., until $\gamma \leqslant 0.90$. The selected values of n may either be such that $\gamma \geqslant 0.90$ or γ is as close as possible to 0.90, as before, but at least here the choice is always between two consecutive numbers for n. From Table 1.1 the answers in these two cases are $n = 15$ and $n = 14$, respectively.

This example shows that the normal test here requires only nine observations to be as powerful as a sign test using 14 or 15, so that the power efficiency is around 0.60 or 0.64. This result applies only for the particular numbers α and β (or γ) selected and therefore is not in any sense a general comparison even though both the null and alternative hypotheses are simple.

Since fixing the value for α is a well-accepted procedure in statistical inference, we might perform calculations similar to those above for some additional and arbitrarily selected values of γ and plot the coordinates (γ, n) and (γ, n^*) on the same graph. From these points the curves $n(\gamma)$ and $n^*(\gamma)$ can be approximated. The numerical processes can be easily programmed for computer calculation. Some evaluation of general relative performance of two tests can therefore be made for the particular value of α selected. However, this power-efficiency approach is satisfactory only for a simple alternative hypothesis. Especially in the case of nonparametric tests, the alternative of interest is usually composite. In the above example, if the alternative were $H_1\colon \mu > 0$ $(M > 0)$, curves for the functions $n[\gamma(\mu)]$ and $n^*[\gamma(\mu)]$ would have to be compared for all $\mu > 0$ and a preselected α. General conclusions for any μ and γ are certainly difficult if not impossible. As a result, we usually make comparisons of the power for μ in a specified neighborhood of the null hypothesis.

In many important cases the limit of the ratio n^*/n turns out not to be a function of α and γ, or even the parameter values when it is in the neighborhood of the hypothesized value. Therefore, even though it is a large-sample property for a limiting type of alternative, the asymptotic relative efficiency of two tests is a somewhat more satisfying criterion for comparison in the sense that it leads to a single number and consequently a well-defined conclusion for large sample sizes. It is for this reason that the discussion here will be limited to comparisons of tests using this standard.

13.2 THEORETICAL BASES FOR CALCULATING THE ARE

Suppose that we have two test statistics T_n and T_n^*, for data consisting of n observations, and both statistics are consistent for a test of

$$H_0: \theta \in \omega \qquad \text{versus} \qquad H_1: \theta \in \Omega - \omega$$

In other words, for all $\theta \in \Omega - \omega$

$$\lim_{n\to\infty} \mathrm{Pw}[T_n(\theta)] = 1 \qquad \text{and} \qquad \lim_{n\to\infty} \mathrm{Pw}[T_n^*(\theta)] = 1$$

Suppose further that a subset of the space Ω can be indexed in terms of a sequence of parameters $\{\theta_0, \theta_1, \theta_2, \ldots, \theta_n, \ldots\}$ such that θ_0 specifies a value in ω and the remaining $\theta_1, \theta_2, \ldots$ are in $\Omega - \omega$ and that $\lim_{n\to\infty} \theta_n = \theta_0$. For example, in the case of a one-sided alternative $\theta > \theta_0$, we take a monotonic decreasing sequence of numbers $\theta_1, \theta_2, \ldots,$ which converges to θ_0 from above. If each θ_i specifies a probability distribution for the test statistics, we might say that the alternative distribution is getting closer and closer to the null distribution as n approaches infinity. Under these conditions, a formal definition of the ARE of T relative to T^* can be given.

Definition *Let* $\mathrm{Pw}_n(\theta)$ *and* $\mathrm{Pw}_n^*(\theta)$ *be the power functions of two tests* T *and* T^* *(corresponding to the test statistics* T_n *and* T_n^**, respectively), against a family of alternatives labeled by* θ*, and let* θ_0 *be the value of* θ *specified by the null hypothesis. Also let* T *and* T^* *have the same level of significance* α*. Consider a sequence of alternatives* $\{\theta_n\}$ *and a sequence* $\{n^*\} = \{h(n)\}$ *of positive integers, where h is some suitable function, such that*

$$\lim_{n\to\infty} \mathrm{Pw}_n(\theta_n) = \lim_{n\to\infty} \mathrm{Pw}_n^*(\theta)$$

where it is assumed that the two limits exist and are not equal to either 0 or 1. Then the asymptotic relative efficiency (ARE) of test T *relative to test* T^* *is*

$$\mathrm{ARE}(T, T^*) = \lim_{n\to\infty} \frac{n^*}{n}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

provided that the limit exists and is independent of the sequences $\{\theta_n\}, \{n\}$, and $\{n^*\}$.

In other words, the ARE is the inverse ratio of the sample sizes necessary to obtain any power γ for the tests T and T^*, respectively, while simultaneously the sample sizes approach infinity and the sequences of alternatives approach θ_0, and both tests have the same significance level. It is thus a measure of asymptotic and localized power efficiency. In the case of the more general tests of hypotheses concerning distributions like $F = F_\theta$, the same definition holds.

Now suppose that our consistent size α tests T_n and T_n^* are for the one-sided alternative

$$H_0\colon \theta = \theta_0 \qquad \text{versus} \qquad H_1\colon \theta > \theta_0$$

and have respective rejection regions of the form

$$T_n \in R \text{ for } T_n \geqslant t_{n,\alpha} \qquad \text{and} \qquad T_n^* \in R^* \text{ for } T_n^* \geqslant t_{n,\alpha}^*$$

where $t_{n,\alpha}$ and $t_{n,\alpha}^*$ are chosen such that

$$P(T_n \geqslant t_{n,\alpha} | \theta = \theta_0) = \alpha \qquad \text{and} \qquad P(T_n^* \geqslant t_{n,\alpha}^* | \theta = \theta_0) = \alpha$$

The following regularity conditions for the test T_n, and analogous ones for T_n^*, must be satisfied.

1. $dE(T_n)/d\theta$ exists and is positive and continuous at θ_0. All other higher-order derivatives, $d^rE(T_n)/d\theta^r$, $r = 2, 3, \ldots,$ are equal to zero at θ_0.
2. There exists a positive constant c such that

$$\lim_{n\to\infty} \frac{dE(T_n)/d\theta|_{\theta=\theta_0}}{\sqrt{n}\sigma(T_n)|_{\theta=\theta_0}} = c$$

3. There exists a sequence of alternatives $\{\theta_n\}$ such that for some constant $d > 0$, we have

$$\theta_n = \theta_0 + \frac{d}{\sqrt{n}}$$

$$\lim_{n\to\infty} \frac{dE(T_n)/d\theta|_{\theta=\theta_n}}{dE(T_n)/d\theta|_{\theta=\theta_0}} = 1$$

$$\lim_{n\to\infty} \frac{\sigma(T_n)|_{\theta=\theta_n}}{\sigma(T_n)|_{\theta=\theta_0}} = 1$$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

4. $\lim_{n\to\infty} P\left[\frac{T_n - E(T_n)|_{\theta=\theta_n}}{\sigma(T_n)|_{\theta=\theta_n}} \leqslant z\,\theta = \theta_n\right] = \Phi(z)$

5. $\lim_{n\to\infty} P[T_n \geqslant t_{n,\alpha}|\theta = \theta_0] = \alpha \qquad 0 < \alpha < 1$

Theorem 2.1 *Under the five regularity conditions above, the limiting power of the test* T_n *is*

$$\lim_{n\to\infty} \mathrm{Pw}(T_n|\theta = \theta_n) = 1 - \Phi(z_\alpha - dc)$$

where z_α *is that number for which* $1 - \Phi(z_\alpha) = \alpha$.

Proof The limiting power is

$$\begin{aligned}
&\lim_{n\to\infty} P(T_n \geqslant t_{n,\alpha}|\theta = \theta_n) \\
&\quad = \lim_{n\to\infty} P\left[\frac{T_n - E(T_n)|_{\theta=\theta_n}}{\sigma(T_n)|_{\theta=\theta_n}} \geqslant \frac{t_{n,\alpha} - E(T_n)|_{\theta=\theta_n}}{\sigma(T_n)|_{\theta=\theta_n}}\right] \\
&\quad = 1 - \Phi(z) \qquad \text{from regularity condition 4}
\end{aligned}$$

where

$$\begin{aligned}
z &= \lim_{n\to\infty} \frac{t_{n,\alpha} - E(T_n)|_{\theta=\theta_n}}{\sigma(T_n)|_{\theta=\theta_n}} \\
&= \lim_{n\to\infty} \left[\frac{t_{n,\alpha} - E(T_n)|_{\theta=\theta_n}}{\sigma(T_n)|_{\theta=\theta_0}} \frac{\sigma(T_n)|_{\theta=\theta_0}}{\sigma(T_n)|_{\theta=\theta_n}}\right] \\
&= \lim_{n\to\infty} \frac{t_{n,\alpha} - E(T_n)|_{\theta=\theta_n}}{\sigma(T_n)|_{\theta=\theta_0}} \qquad \text{from regularity condition 3}
\end{aligned}$$

Expanding $E(T_n)|_{\theta=\theta_n}$ in a Taylor's series about θ_0 and using regularity condition 1, we obtain

$$E(T_n)|_{\theta=\theta_n} = E(T_n)|_{\theta=\theta_0} + (\theta_n - \theta_0)\left.\frac{dE(T_n)}{d\theta}\right|_{\theta=\theta_0^*} \qquad \theta_0 < \theta_0^* < \theta_n$$

Substituting this in the above expression for z, we obtain

$$z = \lim_{n\to\infty} \left\{\frac{t_{n,\alpha} - E(T_n)|_{\theta=\theta_0}}{\sigma(T_n)|_{\theta=\theta_0}} - \frac{(\theta_n - \theta_0)[dE(T_n)/d\theta]|_{\theta=\theta_0^*}}{\sigma(T_n)|_{\theta=\theta_0}}\right\}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$= \lim_{n\to\infty} \left[\frac{t_{n,\alpha} - E(T_n)|_{\theta=\theta_0}}{\sigma(T_n)|_{\theta=\theta_0}} \right] - dc$$

using regularity conditions 1, 2, and 3

$$= z_\alpha - dc$$

Using regularity conditions 5 and 4, we have

$$\alpha = \lim_{n\to\infty} P(T_n \geqslant t_{n,\alpha} | \theta = \theta_0)$$
$$= \lim_{n\to\infty} P\left[\frac{T_n - E(T_n)|_{\theta=\theta_0}}{\sigma(T_n)|_{\theta=\theta_0}} \geqslant \frac{t_{n,\alpha} - E(T_n)|_{\theta=\theta_0}}{\sigma(T_n)|_{\theta=\theta_0}} \right]$$
$$= 1 - \Phi(z_\alpha)$$

This completes the proof.

Theorem 2.2 *If T and T^* are two tests satisfying the regularity conditions above, the* ARE *of T relative to T^* is*

$$\text{ARE}(T, T^*) = \lim_{n\to\infty} \left[\frac{dE(T_n)/d\theta|_{\theta=\theta_0}}{dE(T_n^*)/d\theta|_{\theta=\theta_0}} \right]^2 \frac{\sigma^2(T_n^*)|_{\theta=\theta_0}}{\sigma^2(T_n)|_{\theta=\theta_0}} \tag{2.1}$$

Proof From Theorem 2.1, the limiting powers of tests T and T^*, respectively, are

$$1 - \Phi(z_\alpha - dc) \qquad \text{and} \qquad 1 - \Phi(z_\alpha - d^*c^*)$$

These quantities are equal if

$$z_\alpha - dc = z_\alpha - d^*c^*$$

or, equivalently, for

$$\frac{d^*}{d} = \frac{c}{c^*}$$

From regularity condition 3, the sequences of alternatives are the same if

$$\theta_n = \theta_0 + \frac{d}{\sqrt{n}} = \theta_n^* = \theta_0 + \frac{d^*}{\sqrt{n^*}}$$

or, equivalently, for

$$\frac{d}{\sqrt{n}} = \frac{d^*}{\sqrt{n^*}} \qquad \text{or} \qquad \frac{d^*}{d} = \left(\frac{n^*}{n}\right)^{1/2}$$

Since the ARE is the limit of the ratio of sample sizes when the limiting power and sequence of alternatives are the same for both tests, we have

$$\begin{aligned}
\mathrm{ARE}(T, T^*) &= \frac{n^*}{n} = \left(\frac{d^*}{d}\right)^2 = \left(\frac{c}{c^*}\right)^2 \\
&= \lim_{n\to\infty} \left[\frac{dE(T_n)/d\theta|_{\theta=\theta_0}}{\sqrt{n}\sigma(T_n)|_{\theta=\theta_0}} \frac{\sqrt{n}\sigma(T_n^*)|_{\theta=\theta_0}}{dE(T_n^*)/d\theta|_{\theta=\theta_0}} \right]^2 \\
&= \lim_{n\to\infty} \frac{\left.\frac{[dE(T_n/d\theta]^2}{\sigma^2(T_n)}\right|_{\theta=\theta_0}}{\left.\frac{[dE(T_n^*)/d\theta]^2}{\sigma^2(T_n^*)}\right|_{\theta=\theta_0}}
\end{aligned} \tag{2.2}$$

which is equivalent to (2.1).

From expression (2.2) we see that when these regularity conditions are satisfied, the ARE can be interpreted as the limit an n approaches infinity of the ratio of two quantities

$$\mathrm{ARE}(T, T^*) = \lim_{n\to\infty} \frac{e(T_n)}{e(T_n^*)} \tag{2.3}$$

where $e(T_n)$ is called the *efficacy* of the statistic T_n when used to test the hypothesis $\theta = \theta_0$ and

$$e(T_n) = \frac{[dE(T_n)/d\theta]^2|_{\theta=\theta_0}}{\sigma^2(T_n)|_{\theta=\theta_0}} \tag{2.4}$$

Theorem 2.3 *The statement in Theorem* 2.2 *remains valid as stated if both tests are for a two-sided alternative,* $H_1\colon \theta \neq \theta_0$, *with rejection region*

$$T_n \in R \qquad \text{for } T_n \geqslant t_{n,\alpha_1} \qquad \text{or} \qquad T_n \leqslant t_{n,\alpha_2}$$

where the size of the test is still α, *and a corresponding rejection region is defined for* T_n^* *with the same* α_1 *and* α_2.

Note that the result for the ARE in Theorem 2.2 is independent of both the quantities α and γ. Therefore, when the regularity conditions are satisfied, the ARE does not suffer the disadvantages of the power-efficiency criterion. However, it is only an approximation to relative efficiency for any finite sample size and/or alternative not in the neighborhood of the null case.

In the two-sample cases, where the null hypothesis is equal distributions, if the hypothesis can be parameterized in terms of θ, the same theorems can be used for either one- or two-sided test. The limiting process must be restricted by assuming that as m and n approach infinity, the ratio m/n approaches λ, a constant. When m is approximately a fixed proportion of n regardless of the total sample size $m+n$, the theory goes through as before for n approaching infinity. For two-sample linear rank tests, evaluation of the efficacies is simplified by using the general results for mean and variance given in Theorem 7.3.8.

In various k-sample problems where the null and the alternative hypotheses involve more than one parameter, the result of Theorem 2.2 cannot be directly used to calculate the ARE of one test relative to another. However, the general approach in Theorems 2.1 and 2.2 can be used to derive the ARE in such cases. It may be noted that the theory of asymptotic relative efficiency remains applicable in principle as long as the two competing test statistics have the same form of asymptotic distributions, not necessarily the normal. In this regard it can be shown that when the asymptotic distributions of the test statistics are chi-square distributions, the ARE is equal to the ratio of the noncentrality parameters. For details about these and related interesting results, see, for example, Andrews (1954), Puri (1964), Puri and Sen (1971), and Chakraborti and Desu (1991).

13.3 EXAMPLES OF THE CALCULATION OF EFFICACY AND ARE

We now give some examples of the calculation of efficacy and ARE. In each case the appropriate regularity conditions are satisfied but we do not verify this. This is left as an exercise for the reader.

ONE-SAMPLE AND PAIRED-SAMPLE PROBLEMS

In the one-sample and paired-sample problems treated in Chapter 5, the null hypothesis concerned the value of the population median or median of the population of differences of pairs. This might be called a one-sample location problem with the distribution model.

$$F_X(x) = F(x-\theta) \tag{3.1}$$

for some continous distribution F with median zero. Since F_X then has median θ, the model implies the null hypothesis

$$H_0\colon \theta = 0$$

against one- or two-sided alternatives.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The nonparametric tests for this model can be considered analogs of the one-sample or paired-sample Student's t test for location of the mean or difference of means if F is any continuous distribution symmetric about zero since then θ is both the mean and median of F_X. For a single random sample of size N from any continuous population F_X with mean μ and variance σ^2, the t test statistic of the null hypothesis

$$H_0: \mu = 0$$

is

$$T_N^* = \frac{\sqrt{N}\bar{X}_N}{S_N} = \left[\frac{\sqrt{N}(\bar{X}_N - \mu)}{\sigma} + \frac{\sqrt{N}\mu}{\sigma}\right]\frac{\sigma}{S_N}$$

where $S_N^2 = \sum_{i=1}^N (X_i - \bar{X})^2/(N-1)$. Since $\lim_{N\to\infty}(S_N/\sigma) = 1$, T_N^* is asymptotically equivalent to the Z (normal-theory) test for σ known. The moments of T_N^*, for large N, then are

$$E(T_N^*) = \frac{\sqrt{N}\mu}{\sigma} \qquad \text{and} \qquad \text{var}(T_N^*) = \frac{N\text{var}(\bar{X}_N)}{\sigma^2} = 1$$

and

$$\frac{d}{d\mu}E(T_N^*)|_{\mu=0} = \frac{\sqrt{N}}{\sigma}$$

Using (2.4), the efficacy of Student's t test for observations from any continuous population F_X with mean μ and variance σ^2 is

$$e(T_N^*) = \frac{N}{\sigma^2} \tag{3.2}$$

The ordinary sign-test statistic K_N of Section 5.4 is appropriate for the model (3.1) with

$$H_0: M = \theta = 0$$

Since K_N follows the binomial probability distribution, its moments are

$$E(K_n) = Np \qquad \text{and} \qquad \text{var}(K_N) = Np(1-p)$$

where

$$p = P(X > 0) = 1 - F_X(0)$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

If θ is median of the population F_X, $F_X(0)$ is a function of θ, and for the location model (3.1) we have

$$\begin{aligned}\left.\frac{dp}{d\theta}\right|_{\theta=0} &= \left.\frac{d}{d\theta}[1 - F_X(0)]\right|_{\theta=0}\\ &= \left.\frac{d}{d\theta}[1 - F(-\theta)]\right|_{\theta=0}\\ &= f(\theta)|_{\theta=0} = f(0) = f_X(\theta)\end{aligned}$$

where $\theta = 0$, $p = 0.5$, so that $\operatorname{var}(K_N)|_{\theta=0} = N/4$. The efficacy of the ordinary sign test for N observation from any population F_X with median θ is therefore

$$e(K_N) = 4Nf_X^2(\theta) = 2Nf^2(\theta) = 4Nf^2\left[F^{-1}(0.5)\right] \tag{3.3}$$

We now calculate the efficacy of the Wilcoxon signed-rank test described in Section 5.7. Let $X_1, X_2, \ldots, X_N$ be a random sample from a continuous cdf $F_X(x) = F(x - M)$, where F is symmetrically distributed about 0. Thus the X_is are symmetrically distributed about the median M. The Wilcoxon signed-rank test based on T_N^+ is appropriate to test the null hypothesis H_0: $M = M_0$ where M_0 is specified. However, in order to find the efficacy it will be more convenient to work with $V_N^+ = T_N^+/\binom{N}{2}$. It is clear that a test based on T_N^+ is equivalent to a test based on V_N^+ and hence the two tests have the same efficacy. The mean of V_N^+ is obtained from (5.7.5) as

$$\begin{aligned}E(V_N^+) &= \left(\frac{2}{N-1}\right)P(D_i > 0) + P(D_i + D_j > 0)\\ &= \left(\frac{2}{N-1}\right)[1 - F(-M)] + \int_{-\infty}^{\infty} [1 - F(-x - M)]\, dF(x - M)\\ &= \left(\frac{2}{N-1}\right)[1 - F(-M)] + \int_{-\infty}^{\infty} [1 - F(-y - 2M)]\, dF(y)\end{aligned}$$

Thus, we obtain

$$\frac{dE(V_N^+)}{dM} = \left(\frac{2}{N-1}\right)f(-M) + 2\int_{-\infty}^{\infty} f(-y - 2M)]\, dF(y)$$

after interchanging the order of differentiation and integration. This can be shown to be valid if $f(x) = dF(x)/dx$ is bounded by some positive quantity. Since F is symmetric about 0, $f(y) = f(-y)$, and so

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$\left.\frac{dE(V_N^+)}{dM}\right|_{M=0}. = 2\left[\frac{f(0)}{N-1} + I\right]$$

where

$$I = \int_{-\infty}^{\infty} f(y)\, dF(y) = \int_{-\infty}^{\infty} f^2(y)\, dy$$

Also, from (5.7.6) the variance of V_N^+ under H_0 is

$$\frac{(N+1)(2N+1)}{6N(N-1)^2}$$

Therefore, using (2.4), the efficacy of the Wilcoxon signed-rank test for N observations from a continuous population which is symmetric about θ, is

$$\frac{24[f(0)/(N-1) + I]^2 N(N-1)^2}{(N+1)(2N+1)} \tag{3.4}$$

We can use the efficacy results to calculate the asymptotic relative efficiencies between any two of these tests. For example, from (3.3) and (3.2), the ARE of the sign test relative to the Student's t test is

$$\text{ARE}(K, T^*) = 4\{f[F^{-1}(0.5)]\}^2 \sigma^2 \tag{3.5}$$

The ARE of the Wilcoxon signed-rank test relative to the Student's t test is obtained from (3.4) and (3.2) along with (2.3) as

$$\begin{aligned} &\text{ARE}(T^+, T^*) \\ &\quad = \lim_{N\to\infty} \frac{24[f(0)/(N-1) + I]^2 N(N-1)^2/(N+1)(2N+1)}{N/\sigma^2} \\ &\quad = 12\sigma^2 I^2 \end{aligned} \tag{3.6}$$

The quantity I^2 appears frequently in the ARE expressions of many well-known nonparametric tests. In practice, it may be of interest to estimate I from the sample data in order to estimate the ARE. This interesting problem has been studied by Aubuchon and Hettmansperger (1984).

For the ARE of the sign test relative to the Wilcoxon signed-rank test we obtain, using (3.4) and (3.3) and applying (2.3),

$$\text{ARE}(K, T^+) = \frac{[f\{F^{-1}(0.5)\}]^2}{3[\int_{-\infty}^{\infty} f^2(y)\, dy]^2} \tag{3.7}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

We illustrate the calculations involved by computing the ARE of the sign test relative to the t test, ARE(K, T^*), for the normal, the uniform, and the double exponential distributions, respectively.

1. Normal distribution

$$F_X \text{ is } N(\theta, \sigma^2) \qquad F_X(x) = \Phi\left(\frac{x-\theta}{\sigma}\right) \quad \text{or} \quad F(x) = \Phi\left(\frac{x}{\sigma}\right)$$

$$f(0) = (2\pi\sigma^2)^{-1/2} \qquad e(K_N) = 2N/\pi\sigma^2$$

$$\text{ARE}(K, T^*) = 2/\pi$$

2. Uniform distribution

$$f_X(x) = 1 \qquad \text{for } \theta - 1/2 < x < \theta + 1/2$$

or

pt

$$f(x) = 1 \qquad \text{for } -1/2 < x < 1/2$$

$$f(0) = 1 \qquad \text{var}(X) = 1/12$$

$$e(T_N^*) = 12N \quad e(K_N) = 4N$$

$$\text{ARE}(K, T^*) = 1/3$$

3. Double exponential distribution

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x-\theta|} \quad \text{or} \quad f(x) = \frac{\lambda}{2} e^{-\lambda|x|}$$

$$f(0) = \lambda/2 \qquad \text{var}(X) = 2/\lambda^2$$

$$e(T_N^*) = N\lambda^2/2 \quad e(K_N) = N\lambda^2$$

$$\text{ARE}(K, T^*) = 2$$

In order to facilitate comparisons among the tests, the ARE of the Wilcoxon signed-rank test relative to the t test [ARE(T^+, T^*)] and the ARE of the sign test and the Wilcoxon signed-rank test [ARE(K, T)] are calculated for the uniform, the normal, the logistic, and the double exponential distributions. For the same purpose, the ARE of the sign test relative to the t test is also calculated for the logistic distribution. The ARE values are presented in Table 3.1; verification of these results will be left for the reader.

A closer examination of the ARE values in Table 3.1 reveals some interesting facts. First, from the values in the first column it is evident that the Wilcoxon signed-rank test is a strong competitor to the popular Student's t test when a large sample is available. In particular, for the normal distribution, for which the t test is optimal, very little seems to be lost in terms of efficiency when the Wilcoxon signed-rank test is used instead. Moreover, for distributions with heavier tails than the normal, like the uniform, logistic, and double exponential, the signed-rank test is superior in that the ARE is greater than or equal to 1. In fact, it may be recalled that $\mathrm{ARE}(T^{+}, T^{*})$ is never less than 0.864 for any continuous symmetric distribution (Hodges and Lehmann, 1956).

From the ARE values in the second column of Table 3.1 we see that the sign test is much less efficient than the t test for light to moderately heavy-tailed distributions. In particular, for the normal distribution the sign test is only 64% as efficient as the optimal t test. This poor performance is not entirely unexpected since the simple sign test does not utilize all of the sample information generally available. Interestingly, however, this may lead to its superior performance in the case of a heavy-tailed distribution such as the double exponential. Hodges and Lehman (1956) have shown that $\mathrm{ARE}(K, T^{*}) \geqslant 1/3$ for any continuous unimodal symmetric distribution; the lower bound is achieved for the uniform distribution.

Finally, from the third column of Table 3.1 we see that except for the double exponential distribution, the signed-rank test is more efficient than the sign test.

We summarize by saying that the Wilcoxon signed-rank test is a very viable alternative to the popular Student's t test. The test is appropriate under much milder assumptions about the underlying distribution and it either outperforms or comes very close in performance to the t test, in terms of asymptotic relative efficiency, for many commonly encountered distributions. The sign test is usually less efficient; perhaps it is a popular choice because of its ease of use more than its performance.

Table 3.1 Values of ARE(T^{+}, T^{*}), ARE (K, T^{*}), and ARE(K,T^{+}) for some selected probability distributions

Distribution	$\mathrm{ARE}(T^{+}, T^{*})$	$\mathrm{ARE}(K, T^{*})$	$\mathrm{ARE}(K, T^{+})$
Uniform	1	1/3	1/3
Normal	$3/\pi = 0.955$	$2/\pi = 0.64$	2/3
Logistic	$\pi^2/9 = 1.097$	$\pi^2/12 = 0.82$	3/4
Double exponential	1.5	2	4/3

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For the sampling situation where we have paired-sample data and the hypotheses concern the median or mean difference, all results obtained above are applicable if the random variable X is replaced by the difference variable $D = X - Y$. It should be noted that the parameter σ^2 in (3.2) then denotes the variance of the population of differences,

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\mathrm{cov}(X, Y)$$

and the $f_X(\theta)$ in (3.3) now becomes $f_D(\theta)$.

TWO-SAMPLE LOCATION PROBLEMS

For the general location problem in the case of two independent random samples of sizes m and n, the distribution model is

$$F_Y(x) = F_X(x - \theta) \tag{3.8}$$

and the null hypothesis of identical distributions is

$$H_0\colon \theta = 0$$

The corresponding classical test statistic for populations with a common variance σ^2 is the two-sample Student's t test statistic

$$T^*_{m,n} = \sqrt{\frac{mn}{m+m}}\left(\frac{\bar{Y}_n - \bar{X}_m}{S_{m+n}}\right) = \sqrt{\frac{mn}{m+n}}\left(\frac{\bar{Y}_n - \bar{X}_m - \theta}{\sigma} + \frac{\theta}{\sigma}\right)\frac{\sigma}{S_{m+n}}$$

where

$$S^2_{m+n} = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}{m+n-2}$$

is the pooled estimate of σ^2. Since S_{m+n}/σ approaches 1 as $n \to \infty$, $m/n \to \lambda$, the moments of $T^*_{m,n}$ for n large, $\theta = \mu_Y - \mu_X$, are

$$E(T^*_{m,n}) = \theta\frac{\sqrt{mn/(m+n)}}{\sigma}$$

$$\mathrm{var}(T^*_{m,n}) = \frac{mn}{m+n}\,\frac{\sigma^2/m + \sigma^2/n}{\sigma^2} = 1$$

Therefore

$$\frac{d}{d\theta}E(T^*_{m,n}) = \frac{\sqrt{mn/(m+n)}}{\sigma}$$

and the efficacy of Student's t test for any continuous population is

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$e(T^*_{m,n}) = \frac{mn}{\sigma^2(m+n)} \tag{3.9}$$

For the Mann-Whitney test statistic given in (6.6.2), the mean is

$$E(U_{m,n}) = mnP(Y < X) = mnP(Y - X < 0) = mnp$$

A general expression for p was given in (7.6.3) for any distributions. For the location model in (3.4), this integral becomes

$$p = \int_{-\infty}^{\infty} F_X(x - \theta) f_X(x)\, dx$$

so that

$$\frac{d}{d\theta} E(U_{m,n})\bigg|_{\theta=0} = mn \frac{dp}{d\theta}\bigg|_{\theta=0}. = -mn \int_{-\infty}^{\infty} f_X^2(x)\, dx$$

Under $H_0, p = 0.5$ and the variance is found from (6.6.15) to be

$$\mathrm{var}(U_{m,n}) = \frac{mn(m+n+1)}{12}$$

The efficacy then is

$$e(U_{m,n}) = \frac{12mn \left[\int_{-\infty}^{\infty} f_X^2(x)\, dx\right]^2}{m+n+1} \tag{3.10}$$

Using (3.10) and (3.9) and applying (2.4), we find that the expression for the ARE of the Mann-Whitney test relative to the Student's t test given by the expression for the ARE of the Wilcoxon signed-rank test relative to the t test, with F replaced by F_X. Therefore, as before, the ARE of the Mann-Whitney test does not fall below 0.864 as long as F_X is a continuous cdf. There is, however, one important difference between the one-sample and two-sample cases. In the one-sample case with the Wilcoxon signed-rank test, the underlying F is assumed to be symmetric about 0, but no such assumption is necessary about F_X in the two-sample case with the Mann-Whitney test. Thus, in the two-sample case one can evaluate the ARE expression for an asymmetric distribution like the exponential but that is not allowed in the one-sample case.

Now let us find the efficacy of the median test. Recall that the test is based on $U_{m,n}$, the number of X observations that do not exceed Z, the combined sample median. In order to find the efficacy, we examine the mean of $U_{m,n}$. It can be shown (Mood, 1954) that for large m and n,

$$E(U_{m,n}) = mF_X(c)$$

where c satisfies

$$mF_X(c) + nF_Y(c) = \frac{m+n}{2} \tag{3.11}$$

Now

$$\left.\frac{dE(U_{m,n})}{d\theta}\right|_{\theta=0} = m\left.\frac{dF_X(c)}{dc}\frac{dc}{d\theta}\right|_{\theta=0} = mf_X(c)\left.\frac{dc}{d\theta}\right|_{\theta=0} \tag{3.12}$$

For the location model in (3.8) we have from (3.11)

$$mF_X(c) + nF_X(c-\theta) = \frac{m+n}{2} \tag{3.13}$$

Differentiating (3.13) with respect to θ yields

$$mf_X(c)\frac{dc}{d\theta} + nf_X(c-\theta)\left(\frac{dc}{d\theta} - 1\right) = 0$$

Therefore at $\theta = 0$

$$\left.\frac{dc}{d\theta}\right|_{\theta=0} = \frac{n}{m+n} \tag{3.14}$$

Substituting (3.14) in (3.12), we obtain

$$\left.\frac{dE(U_{M,N})}{d\theta}\right|_{\theta=0} = \frac{mn}{m+n}f_X(c)\Big|_{\theta=0} \tag{3.15}$$

Now from (3.13), when $\theta = 0$, we have

$$mF_X(c) + nF_X(c) = \frac{m+n}{2} \qquad \text{so that } c = F_X^{-1}(0.5)$$

and substitution in (3.15) gives

$$\left.\frac{dE(U_{m,n}}{d\theta}\right|_{\theta=0} = \frac{mn}{m+n}f_X\left[F_X^{-1}(0.5)\right] \tag{3.16}$$

From (6.4.5) the null variance of the median test statistic for large m and n is found to be $mn/4(m+n)$. From (2.4) and (3.16), the efficacy of the median test is then

$$e(U_{m,n}) = 4\left(\frac{mn}{m+n}\right)\left[f_X\{F_X^{-1}(0.5)\}\right]^2 \tag{3.17}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

From (3.10), (3.17) and applying (2.4), we see that the ARE expression for the median test relative to the Mann-Whitney (hence of the Wilcoxon rank sum) test is the same as the ARE expression for the sign test relative to the Wilcoxon signed-rank test given in (3.7) with f replaced by f_X. Hence the efficiency values given in Table 3.1 and the resulting comments made earlier for some specific distributions apply equally to the present case.

The ARE of the Mann-Whitney test relative to Student's t test can be found by evaluating the efficacies in (3.9) and (3.10) for any continuous population F_X with variance σ^2. Since Student's t test is the best test for normal distributions satisfying the general location model, we shall use this as an example. If F_X is $N(\mu_X, \sigma^2)$,

$$\int_{-\infty}^{\infty} f_X^2(x)\,dx = \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1} \exp\left[-\left(\frac{x-\mu_X}{\sigma}\right)^2\right] dx$$

$$= \frac{1}{\sqrt{2\pi\sigma}^2}\sqrt{\frac{1}{2}} = (2\sqrt{\pi}\sigma)^{-1}$$

$$e(T^*_{m,n}) = \frac{mn}{\sigma^2(m+n)} \qquad e(U_{m,n}) = \frac{3mn}{\pi\sigma^2(m+n+1)}$$

$$\text{ARE}(U_{m,n}, T^*_{m,n}) = 3/\pi$$

For the uniform and double exponential distributions, the relative efficiencies are 1 and 3/2, respectively (Problem 13.1).

This evaluation of efficacy of the Mann-Whitney test does not make use of the fact that it can be written as a linear rank statistic. As an illustration of how the general results given in Theorem 7.3.8 simplify the calculation of efficiencies, we shall show that the Terry and van der Waerden tests discussed in Section 8.3 are asymptotically optimum rank tests for normal populations differing only in location.

The weights for the van der Waerden test of (8.3.2) in the notation of Theorem 7.3.8 are

$$a_i = \Phi^{-1}\left(\frac{i}{N+1}\right) = \Phi^{-1}\left(\frac{i}{N}\frac{N}{N+1}\right) = J_N\left(\frac{i}{N}\right)$$

The combined population cdf for the general location model of (3.4) is

$$H(x) = \lambda_N F_X(x) + (1-\lambda_N)F_X(x-\theta)$$

so that

$$J[H(x)] = \lim_{N\to\infty} J_N(H) = \Phi^{-1}[\lambda_N F_X(x) + (1-\lambda_N)F_X(x-\theta)]$$

Applying Theorem 7.3.8 now to this J function, the mean is

$$\mu_N = \int_{-\infty}^{\infty} \Phi^{-1}[\lambda_N F_X(x) + (1-\lambda_N)F_X(x-\theta)] f_X(x)\,dx$$

To evaluate the derivative, we note that since

$$\Phi^{-1}[g(\theta)] = y \qquad \text{if and only if } g(\theta) = \Phi(y)$$

it follows that

$$\frac{d}{d\theta} g(\theta) = \varphi(y)\frac{dy}{d\theta} \qquad \text{or} \qquad \frac{dy}{d\theta} = \frac{d[g(\theta)]/d\theta}{\varphi(y)}$$

Therefore the derivative of μ_N above is

$$\left.\frac{d}{d\theta}\mu_N\right|_{\theta=0} = \int_{-\infty}^{\infty} \frac{-(1-\lambda_N) f_X^2(x)}{\varphi\{\Phi^{-1}[F_X(x)]\}}\,dx \tag{3.18}$$

Now to evaluate the variance when $\theta = 0$, we can use Corollary 7.3.8 to obtain

$$\begin{aligned} N\lambda_N\sigma_N^2 &= (1-\lambda_N)\left\{\int_0^1 [\Phi^{-1}(u)]^2\,du - \left[\int_0^1 \Phi^{-1}(u)\,du\right]^2\right\} \\ &= (1-\lambda_N)\left\{\int_{-\infty}^{\infty} x^2\varphi(x)\,dx - \left[\int_{-\infty}^{\infty} x\varphi(x)\,dx\right]^2\right\} \\ &= 1-\lambda_N \end{aligned}$$

The integral in (3.18) reduces to a simple expression when $F_X(x)$ is $N(\mu_X, \sigma^2)$ since then

$$F_X(x) = \Phi\left(\frac{x-\mu_X}{\sigma}\right) \qquad \text{and} \qquad f_X(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu_X}{\sigma}\right)$$

$$\begin{aligned} \left.\frac{d}{d\theta}\mu_N\right|_{\theta=0} &= -\frac{1-\lambda_N}{\sigma^2}\int_{-\infty}^{\infty} \frac{\varphi^2[(x-\mu_X)/\sigma]}{\varphi[(x-\mu_X)/\sigma]}\,dx \\ &= -\frac{1-\lambda_N}{\sigma}\int_{-\infty}^{\infty} \frac{1}{\sigma}\varphi\left(\frac{x-\mu_X}{\sigma}\right)dx = -\frac{1-\lambda_N}{\sigma} \end{aligned}$$

The efficacy of the van der Waerden X_N test in this normal case is then

$$e(X_N) = \frac{N\lambda_N(1-\lambda_N)}{\sigma^2} = \frac{mn}{N\sigma^2} \tag{3.19}$$

which equals the efficacy of the Student's t test $T^*_{m,n}$ given in (3.9).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

TWO-SAMPLE SCALE PROBLEMS

The general distribution model of the scale problem for two independent random samples is

$$F_Y(x) = F_X(\theta x) \tag{3.20}$$

where we are assuming without loss of generality that the common location is zero. The null hypothesis of identical distributions then is

$$H_0: \theta = 1$$

against either one- or two-sided alternatives. Given two independent random samples of sizes m and n, the analogous parametric test for the scale problem is the statistic

$$T^*_{m,n} = \frac{(n-1)\sum_{i=1}^{m}(X_i - \bar{X})^2}{(m-1)\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Since X and Y are independent and in our model above

$$\text{var}(X) = \theta^2\,\text{var}(Y)$$

the expected value is

$$\begin{aligned} E(T^*_{m,n}) &= E\left[\frac{\sum_{i=1}^{m}(X_i - \bar{X})^2}{m-1}\right]E\left[\frac{n-1}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right] \\ &= (n-1)\text{var}(X)E\left[\frac{1}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right] \\ &= (n-1)\theta^2 E\left[\frac{\text{var}(Y)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right] = (n-1)\theta^2 E\left(\frac{1}{Q}\right) \end{aligned}$$

The probability distribution of Q depends on the particular distribution F_X, but in the normal-theory model, where $F_X(x) = \Phi(x), Q$ has the chi-square distribution with $n-1$ degrees of freedom. Therefore we can evaluate

$$\begin{aligned} E\left(\frac{1}{Q}\right) &= \frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{(n-1)/2}}\int_0^\infty x^{-1}e^{-x/2}x^{[(n-1)/2]-1}\,dx = \frac{\Gamma\left(\frac{n-3}{2}\right)}{2\Gamma\left(\frac{n-1}{2}\right)} \\ &= \frac{1}{n-3} \end{aligned}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

$$E(T^*_{m,n}) = \frac{(n-1)\theta^2}{n-3} \qquad \frac{d}{d\theta}E(T^*_{m,n})\bigg|_{\theta=1} = \frac{2(n-1)}{n-3}$$

In this normal-theory model, under the null hypothesis the distribution of $T^*_{m,n}$ is Snedecor's F with $m-1$ and $n-1$ degrees of freedom. Since the variance of the F distribution with ν_1 and ν_2 degrees of freedom is

$$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2}$$

we have here

$$\text{var}(T^*_{m,n})|_{\theta=1} = \frac{2(n-1)^2(N-4)}{(m-1)(n-5)(n-3)^2}$$

The statistic $T^*_{m,n}$ is the best test for the normal-theory model, and its efficacy for this distribution is

$$e(T^*_{m,n}) = \frac{2(m-1)(n-5)}{N-4} \approx \frac{2mn}{N} = 2N\lambda_N(1-\lambda_N) \tag{3.21}$$

We shall now evaluate the efficacy of the Mood and Freund-Ansari-Bradley-Barton-David-Siegel-Tukey tests by applying the results of Theorem 7.3.8 to the two-sample scale model (3.20), for which

$$H(x) = \lambda_N F_X(x) + (1-\lambda_N)F_X(\theta x)$$

For the Mood test statistic of Section 9.2, we write

$$M'_N = N^{-2}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2 Z_i = N^{-2}M_N$$

so that for M'_N

$$a_i = \left(\frac{i}{N} - \frac{N+1}{2N}\right)^2 = J_N\left(\frac{i}{N}\right)$$

$$\lim_{N\to\infty} J_N(H) = (H-0.5)^2$$

The mean of M'_N then is

$$\mu_N = \int_{-\infty}^{\infty} [\lambda_N F_X(x) + (1-\lambda_N)F_X(\theta) - 0.5]^2 f_X(x)\,dx$$

and, after interchanging the order of differentiation and integration, we have

$$\left.\frac{d\mu_N}{d\theta}\right|_{\theta=1} = 2(1-\lambda_N)\int_{-\infty}^{\infty} [F_X(x) - 0.5]\; x\, f_X^2(x)\,dx$$

and the variance under the null hypothesis is

$$N\lambda_N\sigma_N^2 = (1-\lambda_N)\left\{\int_0^1 (u-0.5)^4\,du - \left[\int_0^1 (u-0.5)^2\,du\right]^2\right\}$$

$$= \frac{1-\lambda_N}{180}$$

so that the efficacy for any continuous distribution F_X with median zero is

$$e(M_N) = 720N\lambda_N(1-\lambda_N)\left\{\int_{-\infty}^{\infty} [F_X(x) - \tfrac{1}{2}] x f_X^2(x)\,dx\right\}^2 \qquad (3.22)$$

In order to compare the Mood statistic with the F test statistic, we shall calculate $e(M_N)$ for the normal-theory model, where $F_X(x) = \Phi(x)$. Then

$$\int_{-\infty}^{\infty}[\Phi(x-0.5)]x\varphi^2(x)dx = \int_{-\infty}^{\infty} x\Phi(x)\varphi^2(x)dx - \frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{\infty} x\varphi(x\sqrt{2})dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{x} x\varphi(t)\varphi^2(x)dt dx$$

$$= \int_{-\infty}^{\infty}\varphi(t)\left(\int_t^{\infty} x\,\frac{1}{2\pi}\,e^{-x^2}dx\right)dt$$

$$= \frac{1}{4\pi}\int_{-\infty}^{\infty}\varphi(t)\,e^{-t^2}dt$$

$$= \frac{1}{4\pi}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\,e^{-3/2t^2}dt$$

$$= (4\pi\sqrt{3})^{-1}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For normal distributions, the result then is

$$e(M_N) = 15N\lambda_N(1-\lambda_N)/\pi^2 \qquad \text{ARE}(M_N, T^*_{m,n}) = 15/2\pi^2$$

Using the same procedure for the tests of Section 9.3, we write the weights for the test A'_N so that $(N+1)A'_N = A_N$, where S_N was given in (9.3.1). The result is

$$a_i = \left|\frac{i}{N+1} - \frac{1}{2}\right| = \frac{N}{N+1}\left|\frac{i}{N} - \frac{1}{2} - \frac{1}{2N}\right| = J_N\left(\frac{i}{N}\right)$$

$$J(H) = |H - 0.5|$$

The mean of A'_N is

$$\mu_N = \int_{-\infty}^{\infty} |\lambda_N F_X(x) + (1-\lambda_N)F_X(\theta x) - 0.5| f_X(x)\, dx$$

and, after interchanging the order of differentiation and integration, we have

$$\begin{aligned}\left.\frac{d\mu_N}{d\theta}\right|_{\theta=1} &= (1-\lambda_N)\int_{-\infty}^{\infty} |x\, f_X(\theta x)| f_X(x)\, dx\Big|_{\theta=1} \\ &= (1-\lambda_N)\int_{-\infty}^{\infty} |x|\, f_X^2(x)\, dx\end{aligned}$$

If $f_X(x)$ is symmetric about its zero median, this reduces to

$$\left.\frac{d\mu_N}{d\theta}\right|_{\theta=1} = 2(1-\lambda_N)\int_0^{\infty} x\, f_X^2(x)\, dx \tag{3.23}$$

The variance when $\theta = 1$ is

$$\begin{aligned}N\lambda_N\sigma_N^2 &= (1-\lambda_N)\left[\int_0^1 |u-0.5|^2\, du - \left(\int_0^1 |u-0.5|\, du\right)^2\right] \\ &= \frac{1-\lambda_N}{48}\end{aligned}$$

For $f(x) = \varphi(x)$, the integral in (3.23) is easily evaluated, and the results are

$$e(A_N) = 12N\lambda_N(1-\lambda_N)/\pi^2$$

$$\text{ARE}(A, T^*) = 6/\pi^2 \qquad \text{ARE}(A, M) = 4/5$$

13.4 SUMMARY

In this chapter we covered the concept of asymptotic relative efficiency of nonparametric tests and showed how to calculate this for some popular tests. The exact power of many nonparametric tests is difficult to find and the ARE becomes a useful tool for comparing two competing tests. Two most common criticisms of the concept of (Pitman-Noether) asymptotic relative efficiency are that (1) the comparison is valid only for large sample sizes and (2) the comparison is "local," valid only in a neighborhood close to the null hypothesis. These criticisms have led to some other criteria in the literature for comparing nonparametric tests. Notable among these is a concept of efficiency due to Bahadur (1960a, 1960b, 1967), often called Bahadur efficiency. For further readings on Bahadur efficiency and an interesting comparison of the sign, Wilcoxon signed-rank, and t tests on the basis of such efficiency, the reader is referred to Klotz (1965). With regard to the first criticism that the ARE is basically a tool for a comparison between tests for large sample sizes, it may be noted that the power efficiency of nonparametric tests, especially for small sample sizes, has been a topic of interest for a long time and a vast amount of work, both analytical and empirical, has been reported in the literature. The reader may refer, for example, to the works of Klotz (1963), Gibbons (1964), Arnold (1965), Randles and Hogg (1973), Randles and Wolfe (1979, p. 116), Blair and Higgins (1980), and Gibbons and Chakraborti (1991), among others, where the powers of some of the tests discussed in this chapter are examined for finite sample sizes. Generally, for moderate sample sizes and common significance levels, it appears that the relative power of nonparametric tests is consistent with the results obtained from the corresponding asymptotic relative efficiency. Nikitin (1995) is a good reference for mathematical details related to asymptotic relative efficiency.

We close with a remark that, as noted earlier, when the ARE between two competing tests is equal to one, a choice between the tests cannot be made from the usual Pitman efficiency point of view. Hodges and Lehmann (1970) have proposed a concept called *deficiency* which is useful in these types of situations.

PROBLEMS

13.1 Use the results of Theorem 7.3.8 to evaluate the efficacy of the two-sample Wilcoxon rank-sum test statistic of (8.2.1) for the location model $F_Y(x) = F_X(x - \theta)$ where:

(*a*) F_X is $N(\mu_X, \sigma^2)$ or $F_X(x) = \Phi[(x - \mu_X)/\sigma]$

(*b*) F_X is uniform, with

$$F_X(x) = \begin{cases} 0 & x \leqslant -1/2 \\ x + 1/2 & -1/2 < x \leqslant 1/2 \\ 1 & x > 1/2 \end{cases}$$

(*c*) F_X is double exponential, with

$$F_X(x) = \begin{cases} (1/2)e^{\lambda x} & x \leqslant 0 \\ 1 - (1/2)e^{-\lambda x} & x > 0 \end{cases}$$

13.2. Calculate the efficacy of the two-sample Student's t test statistic in cases (*b*) and (c) of Problem 13.1.

13.3. Use your answers to Problems 13.1 and 13.2 to verify the following results for the ARE of the Wilcoxon rank-sum (or Mann-Whitney) test to Student's t test:

Normal:	$3/\pi$
Uniform:	1
Double exponential:	3/2

13.4. Calculate the efficacy of the sign test and the Student's t test for the location model $F_X(x) = F(x - \theta)$ where θ is the median of F_X and F is the cdf of the logistic distribution

$$F(x) = (1 - e^{-x})^{-1}$$

13.5. Evaluate the efficiency of the Klotz normal-scores test of (9.5.1) relative to the F test statistic for the normal-theory scale model.

13.6. Evaluate the efficacies of the M_N and A_N test statistics and compare their efficiency for the scale model where, as in Problem 13.1:

(*a*) F_X is uniform.
(*b*) F_X is double exponential.

13.7. Use the result in Problem 13.4 to verify that the ARE of the sign test relative to the Student's t test for the logistic distribution is $\pi^2/12$.

13.8. Verify the following results for the ARE of the sign test relative to the Wilcoxon signed-rank test

Uniform:	1/3
Normal:	2/3
Logistic:	3/4
Double exponential:	4/3

13.9. Suppose there are three test statistics, T_1, T_2, and T_3, each of which can be used to test a null hypothesis H_o against an alternative hypothesis H_1. Show that for any pair of tests, say T_1 and T_3, when the relevant AREs exist,

$$\text{ARE}(T_1, T_3) = \text{ARE}(T_1, T_2)\text{ARE}(T_2, T_3) = [\text{ARE}(T_3, T_1)]^{-1}$$

where we take $1/\infty$ to be 0 and 1/0 to be ∞.

14
Analysis of Count Data

14.1 INTRODUCTION

In this chapter we present several different methods of analyzing count data, that is, data representing the number of observations that have one or more specified characteristics, or that respond in a certain manner to a certain kind of stimulus. The simplest example of count data is that used in the quantile tests and sign tests of Chapter 5. The count data situations in this chapter are more involved.

We start with the basic analysis of data presented in a two-way table with r rows and k columns, called an $r \times k$ contingency table, where the cell counts refer to the number of sample observations that have certain cross characteristics. Here we have a test for the null hypothesis that the characteristics are independent or have no association. We can also calculate the contingency coefficient or the phi coefficient to measure the degree of association or dependence. Then we present some special results for $k \times\ 2$ contingency tables.

Another special case of contingency tables is the 2×2 table with fixed row and column totals and we presents Fisher's exact test for this setting. We also cover McNemar's test for comparing two probabilities of success based on paired or dependent samples. Finally we present some methods for analysis of multinomial data.

14.2 CONTINGENCY TABLES

Suppose we have a random sample of observations, and there are two or more properties or characteristics of interest for each subject in the sample. These properties, say $A, B, C, \ldots$, are called sets or families of attributes. Each of these sets has two or more categories of attributes, say $A_1, A_2, \ldots$, for family A, etc. These attributes may be measurable or not, as long as the categories are clearly defined, mutually exclusive, and exhaustive. Some number N of experimental units are observed, and each unit is classified into exactly one category of each family. The number of units classified into each cross-category constitutes the sample data. Such a layout is called a *contingency table* of order $r_1 \times r_2 \times r_3 \times \cdots$ if there are r_1 categories of family A, r_2 of family B, etc. We are interested in a measure of association between the families, or in testing the null hypothesis that the families are completely independent, or that one particular family is independent of the others. In general, a group of families of events are defined to be completely independent if

$$P(A_i \cap B_j \cap C_k \cap \cdots) = P(A_i)P(B_j)P(C_k) \cdots$$

for all $A_i \subset A$, $B_j \subset B$, $C_k \subset C$, etc. For subgroup independence, say that family A is independent of all others, the requirement is

$$P(A_i \cap B_j \cap C_k \cap \cdots) = P(A_i)P(B_j \cap C_k \cap \cdots)$$

For example, in a public-opinion survey concerning a proposed bond issue, the results of each interview or questionnaire may be classified according to the attributes of gender, education, and opinion. Along with the two categories of gender, we might have three categories of opinion, e.g., favor, oppose, and undecided, and five categories of education according to highest level of formal schooling completed. The data may then be presented in a $2 \times 3 \times 5$ contingency table of 30 cells. A tally is placed in the appropriate cell for each interviewee, and these count data may be used to determine whether gender and educational level have any observable relationship to opinion or to find some relative measure of their association.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

For convenience, we shall restrict our analysis to a two-way classification for two families of attributes, as the extension to higher-way layouts will be obvious. Suppose there are r categories of the type A attribute and k categories of the B attribute, and each of N experimental units is classified into exactly one of the rk cross-categories. In an $r \times k$ layout, the entry in the (i, j) cell, denoted by $X_i j$, is the number of items having the cross-classification $A_i \cap B_j$. The contingency table is written in the following form:

	B family				
A family	B_1	B_2	$\cdots$	B_k	*Row total*
A_1	X_{11}	X_{12}	$\cdots$	X_{1k}	$X_{1.}$
A_2	X_{21}	X_{22}	$\cdots$	X_{2k}	$X_{2.}$
	.	.	$\cdots$	.	.
	.	.	$\cdots$	.	.
	.	.	$\cdots$	.	.
A_r	X_{r1}	X_{r2}	$\cdots$	X_{rk}	$X_{r.}$
Column total	$X_{.1}$	$X_{.2}$	$\cdots$	$X_{.k}$	$X_{..} = N$

The total numbers of items classified respectively into the categories A_i and B_j then are the row and column totals $X_{i.}$ and $X_{.j}$, where

$$X_{i.} = \sum_{j=1}^{k} X_{ij} \qquad \text{and} \qquad X_{.j} = \sum_{i=1}^{r} X_{ij}$$

Without making any additional assumptions, we know that the rk random variables $X_{11}, X_{12}, \ldots, X_{rk}$ have the multinomial probability distribution with parameters

$$\theta_{ij} = P(A_i \cap B_j) \qquad \text{where } \sum_{i=1}^{r} \sum_{j=1}^{k} \theta_{ij} = 1$$

so that the likelihood function of the sample is

$$\prod_{i=1}^{r} \prod_{j=1}^{k} (\theta_{ij})^{x_{ij}}$$

The null hypothesis that the A and B classifications are independent affects only the allowable values of these parameters θ_{ij}.

In view of the definition of independence of families, the hypothesis can be stated simply as

$$H_0: \theta_{ij} = \theta_{i.}\theta_{.j} \qquad \text{for all } i = 1, 2, \ldots, r \quad \text{and} \quad j = 1, 2, \ldots, k$$

where

$$\theta_{i.} = \sum_{j=1}^{k} \theta_{ij} = P(A_i) \qquad \theta_{.j} = \sum_{i=1}^{r} \theta_{ij} = P(B_j)$$

If these $\theta_{i.}$ and $\theta_{.j}$ were all specified under the null hypothesis, this would reduce to an ordinary goodness-of-fit test of a simple hypothesis of the multinomial distribution with rk groups. However, the probability distribution is not completely specified under H_0, since only a particular relation between the parameters need be stated for the independence criterion to be satisfied.

The chi-square goodness-of-fit test for composite hypotheses discussed in Section 4.2 is appropriate here. The unspecified parameters must be estimated by the method of maximum likelihood and the degrees of freedom for the test statistic reduced by the number of independent parameters estimated. The maximum-likelihood estimates of the $(r-1)+(k-1)$ unknown independent parameters are those sample functions that maximize the likelihood function under H_0, which is

$$L(\theta_{1.}, \theta_{2.}, \ldots, \theta_{r.}, \theta_{.1}, \theta_{.2}, \ldots, \theta_{.k}) = \prod_{i=1}^{r}\prod_{j=1}^{k} (\theta_{i.}\theta_{.j})^{x_{ij}} \tag{2.1}$$

subject to the restrictions

$$\sum_{i=1}^{r} \theta_{i.} = \sum_{j=1}^{k} \theta_{.j} = 1$$

The maximum-likelihood estimates of these parameters are easily found to be the corresponding observed proportions, or

$$\hat{\theta}_{i.} = X_{i.}/N \quad \text{and} \quad \hat{\theta}_{.j} = X_{.j}/N \qquad \text{for } i = 1, 2, \ldots, r \text{ and } j = 1, 2, \ldots, k$$

so that the rk estimated cell frequencies under H_0 are

$$N\hat{\theta}_{ij} = X_{i.}X_{.j}/N$$

By the results of Section 4.2, the test statistic then is

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(X_{ij} - X_{i.}X_{.j}/N)^2}{X_{i.}X_{.j}/N} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(NX_{ij} - X_{i.}X_{.j})^2}{NX_{i.}X_{.j}} \tag{2.2}$$

which under H_0 has approximately the chi-square distribution with degrees of freedom $rk - 1 - (r - 1 + k - 1) = (r - 1)(k - 1)$. Since non-independence is reflected by lack of agreement between the observed and expected cell frequencies, the rejection region with significance level α is

$$Q \in R \qquad \text{for } Q \geqslant \chi^2_{(r-1)(k-1),\alpha}$$

As before, if any expected cell frequency is too small, say less than 5, the chi-square approximation is improved by combining cells and reducing the degrees of freedom accordingly.

The extension of this to higher-order contingency tables is obvious. For an $r_1 \times r_2 \times r_3$ table, for example, for the hypothesis of complete independence

$$H_0{:}\,\theta_{ijk} = \theta_{i..}\theta_{.j.}\theta_{..k} \qquad \text{for all } i = 1, 2, \ldots, r_1,\ j = 1, 2, \ldots, r_2,\ k = 1, 2, \ldots, r_3$$

the estimates of expected cell frequencies are

$$N\hat{\theta}_{ijk} = X_{i..}X_{.j.}X_{..k}/N^2$$

and the chi-square test statistic is

$$\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\sum_{k=1}^{r_3} (N^2X_{ijk} - X_{i..}X_{.j.}X_{..k})^2/N^2X_{i..}X_{.j.}X_{..k}$$

with

$$r_1r_2r_3 - 1 - (r_1 - 1 + r_2 - 1 + r_3 - 1) = r_1r_2r_3 - r_1 - r_2 - r_3 + 2$$

degrees of freedom. For the hypothesis that family A is independent of B and C,

$$H_0{:}\,\theta_{ijk} = \theta_{i..}\theta_{.jk}$$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

the estimated expected cell frequencies are

$$N\hat{\theta}_{ijk} = X_{i..}X_{.jk}/N$$

and the chi-square test statistic has

$$r_1r_2r_3 - 1 - (r_1 - 1 + r_2r_3 - 1) = (r_1 - 1)(r_2r_3 - 1)$$

degrees of freedom.

If the experimental situation is such that one set of totals is fixed by the experimenter in advance, say the row totals in an $r \times k$ contingency table, the test statistic for a hypothesis of independence is exactly the same as for completely random totals, although the reasoning is somewhat different. The entries in the ith row of the table constitute a random sample of size $x_{i.}$ from a k-variate multinomial population. For each row then, one of the cell entries is determined by the constant total. One of the observable frequencies is redundant for each row, as is one of the probability parameters $P(A_i \cap B_j)$ for each i. Since

$$P(A_i \cap B_j) = P(A_i)P(B_j \mid A_i)$$

and $P(A_i) = x_{i.}/N$ is now fixed, we shall redefine the relevant parameters as $\theta_{ij} = P(B_j \mid A_i)$, where $\sum_{j=1}^{k} \theta_{ij} = 1$. The dimension of the parameter space is then reduced from $rk - 1$ to $r(k - 1)$. The B family is independent of the A_i category if $\theta_{ij} = P(B_j \mid A_i) = P(B_j) = \theta_j$ for all $j = 1, 2, \ldots, k$, where $\sum_{j=1}^{k} \theta_j = 1$. Under H_0 then, $E(X_{ij}) = x_{i.}\theta_j$, and if the θ_j were specified, the test statistic for independence between B and A_i would be

$$\sum_{j=1}^{k} \frac{(X_{ij} - x_{i.}\theta_j)^2}{x_{i.}\theta_j} \tag{2.3}$$

which is approximately chi square distributed with $k - 1$ degrees of freedom. The B family is completely independent of the A family if $\theta_{ij} = \theta_j, j = 1, 2, \ldots, k$, for all $i = 1, 2, \ldots, r$, so that the null hypothesis can be written as

$$H_0\colon \theta_{1j} = \theta_{2j} = \cdots = \theta_{rj} = \theta_j \qquad \text{for } j = 1, 2, \ldots, k$$

The test statistic for complete independence then is the statistic in (2.3) summed over all $i = 1, 2, \ldots, r$,

$$\sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(X_{ij} - x_{i.}\theta_j)^2}{x_{i.}\theta_j} \tag{2.4}$$

which under H_0 is the sum of r independent chi-square variables, each having $k-1$ degrees of freedom, and therefore has $r(k-1)$ degrees of freedom. Of course, in our case the θ_j are not specified, and so the test statistic is (2.4), with the θ_j replaced by their maximum-likelihood estimates and the degrees of freedom reduced by $k-1$, the number of independent parameters estimated. The likelihood function under H_0 of all N observations with row totals fixed is

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{r} \prod_{j=1}^{k} \theta_j^{x_{ij}} = \prod_{j=1}^{k} \theta_j^{x_{.j}}$$

so that $\hat{\theta}_j = X_{.j}/N$. Substituting this result in (2.4), we find the test criterion unchanged from the previous case with random totals given in (2.2), and the degrees of freedom are $r(k-1)-(k-1) = (r-1)(k-1)$, as before. By similar analysis, it can be shown that the same result holds for fixed column totals or both row and column totals fixed.

THE CONTINGENCY COEFFICIENT

As a measure of the degree of association between families in a contingency table classifying a total of N experimental units, Pearson (1904) proposed the *contingency coefficient* C, defined as

$$C = \left(\frac{Q}{Q+N}\right)^{1/2} \tag{2.5}$$

where Q is the test statistic for the hypothesis of independence. If the families are completely independent, the values of Q and C are both small. Further, increasing values of C imply an increasing degree of association since large values of Q are a result of more significant departures between the observed and expected cell frequencies. Although clearly the value of C cannot exceed 1 for any N, a disadvantage of C as a measure of association is that it cannot attain the value 1, as we now show.

For a two-way contingency table of dimension $r \times k$, the maximum value of C is

$$C_{\max} = \left(\frac{t-1}{t}\right)^{1/2} \qquad \text{where } t = \min(r, k)$$

Without loss of generality, we can assume $r \geqslant k$. Then N must be at least r so that there is one element in each row and each column to avoid any zero denominators in the test statistic. Consider N fixed at its smallest value r, so that $x_{i.} = 1$ for $i = 1, 2, \ldots, r$, and x_{ij} is 0 or 1 for all i, j. The number of cells for which $x_{ij} = 1$ for fixed j is $x_{\cdot j}$, and the number for which $x_{ij} = 0$ is $r - x_{\cdot j}$. The value of Q from (2.2) then is

$$\sum_{j=1}^{k} \frac{(r - x_{\cdot j})(0 - x_{\cdot j}/r)^2 + x_{\cdot j}(1 - x_{\cdot j}/r)^2}{x_{\cdot j}/r}$$

$$= \sum_{j=1}^{k} \frac{x_{\cdot j}(r - x_{\cdot j})[x_j + (r - x_{\cdot j})]}{r x_{\cdot j}} = r(k-1)$$

and the contingency coefficient has the value

$$C = \left[\frac{r(k-1)}{rk - r + r}\right]^{1/2} = \left(\frac{k-1}{k}\right)^{1/2}$$

As a result of this property, contingency coefficients for two different sets of count data are not directly comparable to measure association unless $\min(r, k)$ is the same for both tables. For this reason, some people prefer to use the ratio $C/C_{\max}$ as a measure of association in contingency tables. Another coefficient sometimes used to measure association is the *phi coefficient* defined as

$$\phi = \sqrt{Q/N} \tag{2.6}$$

The sampling distribution of C or ϕ is not known. However, this is of no consequence since C and ϕ are both functions of Q, and a test of significance based on Q would be equivalent to a test of significance based on C^2 or ϕ^2.

Example 2.1 Streissguth et al. (1984) investigated the effect of alcohol and nicotine consumption during pregnancy on the resulting children by examining the children's attention span and reaction time at age four. First, the 452 mothers in the study were classified as shown in Table 2.1 according to their levels of consumption of alcohol and nicotine. Test the null hypothesis of no association between levels of consumption.

Table 2.1 Data for Example 2.1

	Nicotine (mg/day)			
Alcohol (oz./day)	*None*	*1–15*	*16 or more*	*Total*
None	105	7	11	123
0.01–0.10	58	5	13	76
0.11–0.99	84	37	42	163
1.00 or more	57	16	17	90
Total	304	65	83	452

Solution The expected frequencies under the null hypothesis are calculated using the row and column totals in Table 2.1. The results are shown in parentheses in Table 2.2. Note that none of the expected frequencies is small, so there is no need to combine cells. The test statistic is $Q = 42.250$ with 6 degrees of freedom. The P value from Table B of the Appendix is $P < 0.001$ and we condlude that association exists. The value of the contingency coefficient from (2.5) is $C = \sqrt{42.250/494.25} = 0.2924$, and the phi coefficient from (2.6) is $\sqrt{42.250/452} = 0.3057$.

The STATXACT solution is shown below. The results agree with ours. The contingency coefficient is labeled Pearson's CC and the phi coefficient is labeled phi.

```
           ****************************************
             STATXACT SOLUTION TO EXAMPLE 2.1
           ****************************************

CONTINGENCY COEFFICIENTS TO MEASURE ASSOCIATION
Contingency Coefficient estimates based on 452 observations.

Coefficient      Estimate    ASE1       95.00% Confidence Interval
-----------      --------    ----       --------------------------
Phi               0.3057    0.02552     (   0.2557,    0.3558)
Pearson's CC      0.2924    0.03650     (   0.2208,    0.3639)
Sakoda's  CC      0.3581    0.04471     (   0.2705,    0.4457)
Tschuprow's CC    0.1954    0.01089     (   0.1740,    0.2167)
Cramer's  V       0.2162    0.04174     (   0.1344,    0.2980)

      Pearson Chi-Square Statistic =        42.25
Asymptotic p-value:(based on Chi-Square distribution with 6 df )
Pr { Test Statistic .GE. Observed } =      0.0000
```

Table 2.2 Expected frequencies

	Nicotine			
Alcohol	*0*	*1–15*	*16 or more*	*Total*
0	105 (82.7)	7 (17.7)	11 (22.6)	123
0.01–0.10	58 (51.1)	5 (10.9)	13 (14.0)	76
0.11–0.99	84 (109.6)	37 (23.4)	42 (30.0)	163
1.00 or more	57 (60.5)	16 (12.9)	17 (16.5)	90
Total	304	65	83	452

14.3 SOME SPECIAL RESULTS FOR $k \times 2$ CONTINGENCY TABLES

In a $k \times 2$ contingency table, the B family is simply a dichotomy with say success and failure as the two possible outcomes. Then it is a simple algebraic exercise to show that the test statistic for independence can be written in an equivalent form as

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{(X_{ij} - X_{i.}X_{.j}/N)^2}{X_{i.}X_{.j}/N} = \sum_{i=1}^{k} \frac{(Y_i - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})} \tag{3.1}$$

where

$$Y_i = X_{i1} \qquad n_i - Y_i = X_{i2} \qquad \hat{p} = \sum_{i=1}^{k} Y_i/N$$

If B_1 and B_2 are regarded as success and failure, and $A_1, A_2, \ldots, A_k$ are termed sample 1, sample 2, ..., and sample k, we see that the chi-square test statistic in (3.1) is the sum of the squares of k standardized binomial variables with parameter p estimated by its consistent estimator $\hat{p}$. Thus the test based on (3.1) is frequently called the *test for the equality of k proportions*, previously covered in Section 10.8 and illustrated here by Example 3.1.

Example 3.1 A marketing research firm has conducted a survey of businesses of different sizes. Questionnaires were sent to 200 randomly selected businesses of each of three sizes. The data on responses

	Business size		
	Small	*Medium*	*Large*
Response	125	81	40

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

are summarized below. Is there a significant difference in the proportion of nonresponses by small, medium, and large businesses?

Solution The frequencies of nonresponses are 75, 119, and 160. The best estimate of the common probability of nonresponse is $(75+119+160)/600=0.59$. The expected numbers of nonresponse are then 118 for each size business. The value of Q from (3.1) is 74.70 with 2 degrees of freedom. From Table B we find $P<0.001$, and we conclude that the proportions of nonresponse are not the same for the three sizes of businesses.

If $k=2$, the expression in (3.1) can be written as

$$Q=\frac{(Y_1/n_1-Y_2/n_2)^2}{\hat{p}(1-\hat{p})(1/n_1+1/n_2)} \tag{3.2}$$

Now the chi-square test statistic in (3.2) is the square of the difference between two sample proportions divided by the estimated variance of their difference. In other words, Q is the square of the classical standard normal theory test statistic used for the hypothesis that two population proportions are equal.

Substituting the original X_{ij} notation in (3.2), a little algebraic manipulation gives another equivalent form for Q as

$$Q=\frac{N(X_{11}X_{22}-X_{12}X_{21})^2}{X_{.1}X_{.2}X_{1.}X_{2.}} \tag{3.3}$$

This expression is related to the sample Kendall tau coefficient of Chapter 11. Suppose that the two families A and B are factors or qualities, both dichotomized into categories which can be called presence and absence of the factor or possessing and not possessing the quality. Suppose further that we have a single sample of size N, and that we make two observations on each element in the sample, one for each of the two factors. We record the observations using the code 1 for presence and 2 for absence. The observations then consist of N sets of pairs, for which the Kendall tau coefficient T of Chapter 11 can be determined as a measure of association between the factors. The numerator of T is the number of sets of pairs of observations, say (a_ib_i) and (a_jb_j), whose differences a_i-a_j and b_i-b_j have the same sign but are not zero. The differences here are both positive or both negative only for a set (1,1) and (2,2), and are of opposite signs for a set (1,2) and (2,1). If X_{ij} denotes the number of observations where factor A was recorded as i and factor B was recorded as j for $i,j=1,2$, the number of differences with the same sign is the product $X_{11}X_{22}$, the number of pairs which agreed in the sense that both factors were present or both were absent. The number of differences with opposite signs is $X_{12}X_{21}$, the number of pairs which disagreed. Since there are so many ties, it seems most appropriate to use the definition of T modified for ties, given in (11.2.37) and called tau b. Then the denominator of T is the square root of the product of the numbers of pairs with no ties for each factor, or $X_{1.}X_{2.}X_{.1}X_{.2}$. Therefore the tau coefficient is

$$T = \frac{X_{11}X_{22} - X_{12}X_{21}}{(X_{.1}X_{.2}X_{1.}X_{2.})^{1/2}} = \left(\frac{Q}{N}\right)^{1/2} \tag{3.4}$$

and Q/N estimates τ^2, the parameter of association between factors A and B. For this type of data, the Kendall measure of association is sometimes called the phi coefficient, as defined in (2.6).

Example 3.2 The researchers in the study reported in Example 2.1 really might have been more interested in a one-sided alternative of positive dependence between the variables alcohol and nicotine. Since the data are measurements of level of consumption, we could regard them as 452 pairs of measurements with many ties. For example, the 37 mothers in cell (3,2) of Table 2.1 represent the pair of measurements (AIII, BII), where AIII indicates alcohol consumption in the 0.11–0.99 range and BII represents nicotine consumption at level 1–15. For these kinds of data we can then calculate Kendall's tau for the 452 pairs. The number of concordant pairs C and the number of discordant pairs Q are calculated as shown in Table 2.3. Because the ties are quite extensive, we need to incorporate the correction for ties in the calculation of T from (11.2.38). Then we use the normal approximation to the distribution of T in (11.2.30) to calculate the right-tailed P value for this one-sided alternative.

Table 2.3 Calculations for C and Q

C	Q
$105(5+13+37+42+16+17) = 13{,}650$	$7(58+84+57) = 1{,}393$
$7(13+42+17) = 504$	$11(58+84+57+5+37+16) = 2{,}827$
$58(37+42+16+17) = 6{,}496$	$58(84+57) = 705$
$5(42+17) = 295$	$13(84+57+37+16) = 2{,}522$
$84(16+17) = 2{,}2772$	$37(57) = 2{,}109$
$37(17) = 629$	$42(57+16) = 3{,}066$
24,346	12,622

$$T = \frac{24{,}346 - 12{,}622}{\sqrt{\left[\binom{452}{2} - \binom{304}{2} - \binom{65}{2} - \binom{83}{2}\right]\left[\binom{452}{2} - \binom{123}{2} - \binom{76}{2} - \binom{163}{2} - \binom{90}{2}\right]}}$$
$$= 0.1915$$

$$Z = \frac{3(0.1915)\sqrt{452(451)}}{\sqrt{2(904+5)}} = 6.08$$

We find P=0.000 from Table A of the Appendix.

There is also a relationship between the value of the chi-square statistic in a 2×2 contingency table and Kendall's partial tau coefficient. If we compare the expression for $T_{XY.Z}$ in (12.6.1) with the expression for Q in (3.3), we see that

$$T_{XY.Z} = \sqrt{Q/N} \quad \text{for } N = \binom{m}{2}$$

A test for the significance of $T_{XY.Z}$ cannot be carried out by using Q, however. The contingency table entries in Table 6.1 of Chapter 12 are not independent even if X and Y are independent for fixed Z, since all categories involve pairings with the Z sample.

14.4 FISHER'S EXACT TEST

Suppose we have two independent samples of sizes n_1 and n_2, from two binomial populations, 1 and 2, with probability of success θ_1 and θ_2, respectively, and observed number of successes y_1 and y_2, respectively. The data can be represented in a 2×2 table as in Table 4.1. The row totals in this table are fixed since they are the designated sample sizes. As discussed in Section 14.3, the Q statistic in (3.2) can be used as an approximate test of the null hypothesis that the success probabilities are equal when the sample sizes are large.

Table 4.1 Presentation of data

	Subject's identification		
Population	*Success*	*Failure*	*Total*
1	y_1	$n_1 - y_1$	n_1
2	y_2	$n_2 - y_2$	n_2
Total	$y_1 + y_2$	$N - (y_1 + y_2)$	N

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

We now present an exact test for this problem that can be used for any sample sizes when the marginal column totals $Y = Y_1 + Y_2$ and therefore also $N - (Y_1 + Y_2)$ are assumed fixed. This is known as *Fisher's exact test*. Fisher's example of applications is where an experiment is designed to test a human's ability to identify (discriminate) correctly between two objects, success and failure, when the subject is told in advance exactly how many successes are in the two samples combined. The subject's job is simply to allocate the total number of successes between the two groups. Under the null hypothesis, this allocation is a random assignment; i.e., the subject is merely guessing.

Note that in the 2×2 table, the marginal row totals are fixed at the two given sample sizes. For a given value of $y_1 + y_2$, the value of y_1 determines the remaining three cell counts. Under the null hypothesis H_0: $\theta_1 = \theta_2 = \theta$, the conditional distribution of Y_1 given the marginal totals is the hypergeometric distribution

$$\frac{\binom{n_1}{y_1}\binom{n_2}{y_2}}{\binom{N}{y}} \tag{4.1}$$

where y is the sum of the values observed under the first column. Inferences can be based on an exact P value calculated from (4.1) for an observed y_1. The premise here is that the observed 2×2 table is one of the many possible 2×2 tables that could have been observed with the row and the column totals fixed at their presently observed values. The question then becomes how extreme the currently observed table (value y_1) is, in the appropriate direction, among all of the possible tables with the same marginal totals. The more extreme it is, the more is the evidence against the null hypothesis.

For example, if the alternative hypothesis is H_1: $\theta_1 > \theta_2$, the null hypothesis should be rejected if Y_1 is large. The exact P value can be calculated by finding the probability $P(Y_1 \geqslant y_{1O} \mid Y = y)$, where y_{1O} is the observed value of Y_1. Again, this P value is calculated from all possible 2×2 tables with the same marginal totals as the observed one, but having a value of Y_1 as extreme as or more extreme than the value y_{1O} of Y_1 for the observed table. We illustrate this test with the famous data from Fisher's tea testing experiment.

Example 4.1 Sir R. A. Fisher, the English statistician, has been called the father of modern statistics. A famous story is that a colleague of Fisher's claimed that she could tell, while drinking tea with milk, whether milk or tea was poured into the cup first. An experiment was designed to test her claim. Eight cups of tea were presented to her in a random order; four of these had milk poured first while the other four had tea poured first. She was told that there were four cups of each type. The following data show the results of the experiment. She was right 3 out of 4 times on both types. Is this sufficient evidence of her claim of special power?

	Guess poured first		
Poured first	*Milk*	*Tea*	*Total*
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Solution The potential values of Y_1 are (0,1,2,3,4) and the observed value is 3, the number of cups with milk poured first that were correctly guessed. Only one other 2×2 table with the same marginal totals is more extreme than the observed table, and this is shown below.

	Guess poured first		
Poured first	*Milk*	*Tea*	*Total*
Milk	4	0	4
Tea	0	4	4
Total	4	4	8

The exact P value is then the sum of the conditional probabilities for these two results calculated from (4.1) or

$$\left[\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}\right] \Big/ \binom{8}{4} = 0.2286 + 0.0143 = 0.2429$$

Hence there is not sufficient evidence to suggest that Fisher's colleague has any special power to determine whether tea or milk was poured into the cup first. The value of the chi-square statistic calculated from (3.3) is $Q = 2.0$ with df $= 1$. The P value from Table B is $0.10 < P < 0.25$ but this is for a two-sided alternative. For a 2×2 table with such small frequencies and a one-sided alternative, the chi-square approximation should be suspect.

The STATXACT and the SAS outputs for this example are shown below. Both show the exact P value for a one-sided test as 0.2429, which agrees with ours. Note that the probability that Y_1 equals 3 (0.2286) also appears on both printouts, but STATXACT labels it as the value of the test statistic. The Fisher statistic in the STATXACT printout (1.807) is not the same as ours and should not be interpreted as such.

```
          ***************************************
           STATXACT SOLUTION TO EXAMPLE 4.1
          ***************************************

FISHER'S EXACT TEST

Statistic based on the observed 2 by 2 table(x) :
   P(X)  = Hypergeometric Prob. of the table =      0.2286
   FI(X) = Fisher statistic                  =      1.807

Asymptotic p-value: (based on Chi-Square distribution with 1 df )
     Two-sided:Pr{FI(X) .GE.        1.807} =      0.1789
     One-sided:0.5 * Two-sided             =      0.0894

Exact p-value and point probabilities :
   Two-sided:Pr{FI(X).GE. 1.807}= Pr{P(X).LE. 0.2286} = 0.4857
             Pr{FI(X).EQ. 1.807}= Pr{P(X).EQ. 0.2286} = 0.4571
 One-sided:Let y be the value in Row 1 and Column 1
   y =3 min(Y) =0 max(Y) =4 mean(Y) = 2.000 std(Y) =      0.7559
           Pr { Y .GE. 3 } =      0.2429
           Pr { Y .EQ. 3 } =      0.2286

          *******************************
           SAS SOLUTION TO EXAMPLE 4.1
          *******************************

Program Code:

DATA TEATEST;
INPUT GROUP $ OUTCOME $ COUNT;
DATALINES;
MILK MILK 3
MILK TEA 1
TEA TEA 3
TEA MILK 1
;
PROC FREQ DATA=TEATEST;
  TABLES GROUP * OUTCOME / FISHER;
  WEIGHT COUNT;
RUN;
```

```
Output:
          The FREQ Procedure

              Table of GROUP by OUTCOME

         GROUP     OUTCOME

         Frequency|
         Percent  |
         Row Pct  |
         Col Pct  |MILK    |TEA     |  Total
         ---------+--------+--------+
         MILK     |     3  |     1  |      4
                  | 37.50  | 12.50  |  50.00
                  | 75.00  | 25.00  |
                  | 75.00  | 25.00  |
         ---------+--------+--------+
         TEA      |     1  |     3  |      4
                  | 12.50  | 37.50  |  50.00
                  | 25.00  | 75.00  |
                  | 25.00  | 75.00  |
         ---------+--------+--------+
         Total          4        4        8
                    50.00    50.00   100.00

       Statistics for Table of GROUP by OUTCOME

Statistic                       DF      Value      Prob
--------------------------------------------------------
Chi-Square                       1     2.0000    0.1573
Likelihood Ratio Chi-Square      1     2.0930    0.1480
Continuity Adj. Chi-Square       1     0.5000    0.4795
Mantel-Haenszel Chi-Square       1     1.7500    0.1859
Phi Coefficient                        0.5000
Contingency Coefficient                0.4472
Cramer's V                             0.5000

WARNING: 100% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.

          Fisher's Exact Test
          -----------------------------------
          Cell (1,1) Frequency (F)          3
          Left-sided Pr <= F           0.9857
          Right-sided Pr >= F          0.2429
          Table Probability (P)        0.2286
          Two-sided Pr <= P            0.4857
          Sample Size = 8
```

```
***************************************************
STATXACT SOLUTION TO POWER OF FISHER'S EXACT TEST
***************************************************

>>> Power: Two Binomials Output

Exact Power of Two-Sided Tests for Comparing Two Binomial Populations

Type I error (Alpha)      = 0.05

Pop 2 probability: Difference of proportions model (Difference = 0.300000)

Type of test: Fisher's exact test

Probabilities (Pi):       0.5     0.8
Sample Size (n):          10      10

Power         = 13%
```

Fig. 4.1 STATXACT output for the power of Fisher's exact test.

We note that the two-sample median test presented in Section 6.4 can be viewed as a special case of Fisher's exact test where the value of $y_1 + y_2$ is the number of observations smaller than the sample median of the two combined samples, which is fixed at $N/2$ if N is even and $(N-1)/2$ if N is odd.

For further discussions on Fisher's exact test, the reader is referred to the review article by Gibbons (1982). STATXACT can calculate the power of Fisher's exact test for a given $\alpha, n_1, n_2, \theta_1 = p_1$ and $\theta_2 = p_2$. For illustration, suppose $n_1 = n_2 = 10$ and α=0.05, and let $p_1 = 0.5$ and $p_2 = 0.8$. The exact power of Fisher's exact test is 0.13, as shown in the output labeled Figure 4.1. The latest version of STATXACT also has options for calculating the sample size for a given α, p_1, p_2 and power.

In the next section we consider the problem of comparing the probabilities of success for two groups with paired or dependent samples.

14.5 McNEMAR'S TEST

Suppose that a 2×2 table of data arises when a success or failure response is observed on each of N subjects before and after some treatment. The paired data are dependent within a pair but independent across pairs. Let X_{11} be the number of subjects whose responses are successes both before and after the treatment and let X_{22} be the number of subjects whose responses are failures both before and after the treatment. Then X_{12} and X_{21} denote the numbers of reversals (or discordant pairs) in responses; that is, X_{12} is the number of subjects whose initial (before treatment) response was success but became failure after the treatment, and similarly for X_{21}. The data can then be summarized in the following 2×2 table.

	After treatment		
Before treatment	*Success*	*Failure*	*Total*
Sucess	X_{11}	X_{12}	$X_{1.}$
Failure	X_{21}	X_{22}	$X_{2.}$
Total	$X_{.1}$	$X_{.2}$	N

The two groups of interest are the subjects before and after the treatment, and the hypothesis of interest is that the probability of success is the same before and after the treatment. Let $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ denote the four cell probabilities for the table, with $\theta_{11} + \theta_{12} + \theta_{21} + \theta_{22} = 1$. The interpretation is that θ_{11} say is the probability of success before and after treatment. The sum $\theta_{.1} = \theta_{11} + \theta_{21}$ is the marginal probability of success after treatment and $\theta_{1.} = \theta_{11} + \theta_{12}$ is the marginal probability of success before treatment. The null hypothesis is then parameterized as H_0: $\theta_{1.} = \theta_{.1}$ but this is the same as $\theta_{12} = \theta_{21}$. In other words the null hypothesis can be viewed as testing that the probability of a reversal in either direction is the same.

For the null hypothesis H_0: $\theta_{1.} = \theta_{.1}$ it is natural to consider a test based on $T = (X_{1.} - X_{.1})/N$, an unbiased estimator of the difference $\theta_{1.} - \theta_{.1}$. Since $X_{1.} = X_{11} + X_{12}$ and $X_{.1} = X_{11} + X_{21}$, T reduces to $T = (X_{12} - X_{21})/N$, the difference between the proportions of discordant pairs (numbers in the off-diagonal positions divided by N). Under the null hypothesis, the mean of T is zero and the variance of T can be shown to be $(\theta_{12} + \theta_{21})/N$. *McNemar's test* for H_0 against the two-sided alternative H_1: $\theta_{1.} \neq \theta_{.1}$ is based on

$$(X_{12} - X_{21})^2/(X_{12} + X_{21}) \tag{5.1}$$

which is approximately distributed as a chi-square with 1 degree of freedom. The reader is warned about the inaccuracy of the chi-square approximation for small expected cell frequencies.

We now derive the variance of $T = (X_{12} - X_{21})/N$. The distributions of X_{12} and X_{21} are each binomial with parameters N, θ_{12} and θ_{21}, respectively. Hence $E(X_{12}) = N\theta_{12}, \mathrm{var}(X_{12}) = N\theta_{12}(1 - \theta_{12})$ and

$E(X_{21}) = N\theta_{21}, \text{var}(X_{21}) = N\theta_{21}(1-\theta_{21})$. This gives $E(T) = \theta_{12} - \theta_{21}$. The variance of T will be found from

$$N^2 \text{var}(T) = \text{var}(X_{12}) + \text{var}(X_{21}) - 2\,\text{cov}(X_{12}, X_{21}) \tag{5.2}$$

In order to find the covariance term, we note that the joint distribution of the counts $(X_{11}, X_{12}, X_{21}, X_{22})$ is a multinomial distribution with probabilities $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$. From this it follows that the joint distribution of X_{12} and X_{21} is a multinomial with probabilities $(\theta_{12}, \theta_{21})$. The moment generating function of X_{12} and X_{21} was given in Table 1.2.1 as

$$\{\theta_{12}e^{t_1} + \theta_{21}e^{t_2} + [1 - (\theta_{12} + \theta_{21})]\}^N \tag{5.3}$$

The reader can verify that by taking the second partial derivative of (5.3) with respect to t_1 and t_2 and setting $t_1 = t_2 = 0$, we obtain the second joint moment about the origin as

$$E(X_{12}X_{21}) = N(N-1)\theta_{12}\theta_{21}$$

Hence the covariance is

$$\text{cov}(X_{12}, X_{21}) = N(N-1)\theta_{12}\theta_{21} - (N\theta_{12})(N\theta_{21}) = -N\theta_{12}\theta_{21}$$

Now substituting back in (5.2) gives

$$\begin{aligned} N^2 \text{var}(T) &= N\theta_{12}(1-\theta_{12}) + N\theta_{21}(1-\theta_{21}) - 2(-N\theta_{12}\theta_{21}) \\ &= N[(\theta_{12}+\theta_{21}) - (\theta_{12}-\theta_{21})^2] \end{aligned} \tag{5.4}$$

Therefore $T = (X_{12} - X_{21})/N$ has expectation $\theta_{12} - \theta_{21}$ and variance

$$[(\theta_{12}+\theta_{21}) - (\theta_{12}-\theta_{21})^2]/N$$

Under the null hypothesis $\theta_{12} = \theta_{21}$, T has zero expectation and variance $(\theta_{12}+\theta_{21})/N$, which can be consistently estimated by $(X_{12}+X_{21})/N^2$. McNemar's test statistic in (5.1) is the square of T divided by this estimated variance.

A second motivation of McNemar's test can be given as follows. As noted before, the joint distribution of the counts $(X_{11}, X_{12}, X_{21}, X_{22})$ is a multinomial distribution with probabilities $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$. Let $S = X_{12} + X_{21}$ denote the total number of discordant pairs. The reader can verify that the conditional distribution of X_{12} given S is binomial $[S, p = \theta_{12}/(\theta_{12}+\theta_{21})]$. Then, since $\theta_{12} = \theta_{21}$ under the null hypothesis, an exact (conditional) P value can be calculated from the binomial $(S, 0.5)$ distribution as the probability that X_{12} is as extreme as or more extreme than its observed value, in the direction of the alternative.

We are usually interested in a one-sided alternative that the treatment is effective, that is H_1: $\theta_{.1} > \theta_{1.}$, and this is equivalent to H_1: $\theta_{12} < \theta_{21}$. For this alternative, the exact P value is given by $\sum_{j=0}^{x_{12}} \binom{s}{j} (0.5)^s$, which can be found from Table G of the Appendix for $s \leqslant 20$. For large sample sizes, an approximate P value can be based on the normal approximation to the binomial distribution with the statistic

$$Z = \frac{X_{12} - 0.5S}{\sqrt{0.25S}} = \frac{X_{12} - X_{21}}{\sqrt{X_{12} + X_{21}}} \tag{5.5}$$

which is approximately standard normal, or, when using a continuity correction,

$$Z = \frac{X_{12} - X_{21} + 0.5}{\sqrt{X_{12} + X_{21}}} \tag{5.6}$$

The squared value of (5.2) is the McNemar test statistic for matched pairs in (5.1), which is approximately distributed as a chi square with one degree of freedom.

For a one-sided alternative H_1: $\theta_{12} > \theta_{21}$, the appropriate rejection region is in the right-tail of Z in (5.5). STATXACT provides both exact and approximate tests, as we show for Example 5.1.

Example 5.1 Suppose that a random sample of 100 subjects is evaluated for pain. Under a placebo, 35 of the subjects reported some relief but the remaining 65 did not. Under a new drug, 55 of the subjects reported some relief while 45 did not. Of the 65 people who reported no relief under the placebo, 30 also reported no relief under the new drug. Is there any evidence that the new drug is more effective from the placebo?

Solution The data can be represented in the following 2×2 table. The alternative of interest is $\theta_{.1} > \theta_{1.}$ or $\theta_{21} > \theta_{12.}$

	New drug		
Placebo	*Some relief*	*No relief*	*Total*
Some relief	20	15	35
No relief	35	30	65
Total	55	45	100

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

The value of the test statistic for a one-sided test is 15. The normal approximation to the binomial probability of a value less than or equal to 15, with a total of 50 observations, is calculated from (5.5) as

$$Z = \frac{15 - 50(0.5)}{\sqrt{0.25(50)}} = -2.83$$

and the P value from Table A of the Appendix is estimated to be 0.0023. The corresponding approximation with a continuity correction from (5.3) is

$$Z = \frac{(15 - 35) + 0.5}{\sqrt{15 + 35}} = -2.76$$

with a P value of 0.0029. In each case, there is sufficient evidence that the new drug is more effective than the placebo.

The value of McNemar's test statistic for a two-sided alternative from (5.1) is

$$Z^2 = \frac{(35 - 15)^2}{35 + 15} = 8.0$$

Table B shows $0.001 < P < 0.005$.

The STATXACT and SAS outputs for this example are shown below. In STATXACT, note that the value of the Z statistic is calculated without the continuity correction. The approximate P values are the same as ours and are fairly close to the exact values. The reader is referred to the STATXACT user's manual for details regarding the exact P value calculations. The value of McNemar's statistic is shown on the SAS output and it agrees with ours. The approximate P value is 0.0047, which is simply twice that of the one-tailed P value associated with the Z statistic.

```
****************************************
   STATXACT SOLUTION TO EXAMPLE 5.1
****************************************

MARGINAL HOMOGENEITY TEST for ordered table

Statistic based on the observed 5 by 5 table(x) with 50 observations:
  Min      Max      Mean     Std-dev   Observed   Standardized
 50.00    100.0    75.00     3.536      65.00          -2.828

Asymptotic Inference:
 One-sided p-value: Pr { Test Statistic .LE. Observed } = 0.0023
 Two-sided p-value: 2 * One-sided                       = 0.0047

Exact Inference:
 One-sided p-value: Pr { Test Statistic .LE. Observed } = 0.0033
                    Pr { Test Statistic .EQ. Observed } = 0.0020
 Two-sided p-value: 2*One-Sided                         = 0.0066
```


Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

```
**********************************
     SAS SOLUTION TO EXAMPLE 5.1
**********************************
DATA PAIN;
INPUT DRUG $ PLACEBO $ COUNT;
DATALINES;
YES YES 20
YES NO 35
NO YES 15
NO NO 30
;
PROC FREQ DATA=PAIN;
   TABLES PLACEBO * DRUG / AGREE;
   WEIGHT COUNT;
RUN;
```

Output:

```
                 The FREQ Procedure
              Table of PLACEBO by DRUG
PLACEBO     DRUG
Frequency|
Percent  |
Row Pct  |
Col Pct  |NO      |YES     |  Total
---------+--------+--------+
NO       |     30 |     35 |     65
         |  30.00 |  35.00 |  65.00
         |  46.15 |  53.85 |
         |  66.67 |  63.64 |
---------+--------+--------+
YES      |     15 |     20 |     35
         |  15.00 |  20.00 |  35.00
         |  42.86 |  57.14 |
         |  33.33 |  36.36 |
---------+--------+--------+
Total          45       55      100
            45.00    55.00   100.00

        Statistics for Table of PLACEBO by DRUG
                  McNemar's Test
             ---------------------------
             Statistic (S)    8.0000
             DF                    1
             Pr > S           0.0047

             Simple Kappa Coefficient
         ------------------------------------
         Kappa                        0.0291
         ASE                          0.0919
         95% Lower Conf Limit        -0.1511
         95% Upper Conf Limit         0.2093
                 Sample Size = 100
```


MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

The power of McNemar's test has been studied by various authors. A related issue is the determination of sample size. The reader is referred to Lachin (1993) and Lachenbruch (1992) and the references in these papers. The latest version of STATXACT has options for calculating the power and the sample size for McNemar's test.

14.6 ANALYSIS OF MULTINOMIAL DATA

Count data can also arise when sampling from a multinomial distribution where we have k possible categories or outcomes with respective probabilities $p_1, p_2, \ldots, p_k$, which sum to one. We can use the chi-square goodness-of-fit in Section 4.2 to test the null hypothesis that the sample data conform to specified values for these probabilities (see Problems 4.1, 4.3 to 4.5, 4.27, 4.31, and 4.32).

If we have random samples from two or more multinomial distributions, each with the same k possible categories or outcomes, the data can be presented in an $r \times k$ contingency table where the rows represent the samples and the columns represent the categories. Now X_{ij} denotes the number of outcomes in category j for the ith sample, and the probabilities of these outcomes for the ith sample are denoted by $p_{i1}, p_{i2}, \ldots, p_{ik}$ where $0 < p_{ij} < 1$ and $\sum_j p_{ij} = 1$. We will consider only the case where we have $r = 2$ samples of sizes n_1 and n_2. The data can be presented in a $2 \times k$ table as in Table 6.1. Note that the row totals are fixed by the sample sizes.

We are interested in testing the null hypothesis H_0: $p_{11} = p_{21}$, $p_{12} = p_{22}, \ldots, p_{1k} = p_{2k}$. The common probability for the jth category is estimated by $(X_{1j} + X_{2j})/N = X_{.j}/N$, and the estimated cell frequencies are $n_1 X_{.j}/N$ and $n_2 X_{.j}/N$ for samples 1 and 2, respectively. The chi-square test statistic with $df = k - 1$ is then

$$Q = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(X_{ij} - n_i X_{.j}/N)^2}{n_i X_{.j}/N} \tag{6.1}$$

Table 6.1 Presentation of data

	Category or outcome				
Sample	*1*	*2*	...	*k*	*Total*
1	X_{11}	X_{12}	...	X_{1k}	n_1
2	X_{21}	X_{22}	...	X_{2k}	n_2
Total	$X_{.1}$	$X_{.2}$	...	$X_{.k}$	N

which is the same as (3.1), the test for equality of k proportions.

Example 6.1 Businesses want to maximize return on any money spent on advertising. If the medium is a television commercial, they want the audience to remember the main points of the commercial as long as possible. Two versions of a commercial were test marketed on 100 volunteers. The volunteers were randomly assigned to two groups to view commercials A or B so that each group had 50 volunteers. After 2 days the participants were telephoned and asked to classify their recollection of the commercial as either "Don't remember," "Remember vaguely," or "Remember key points." The data are shown below. Are commercials A and B equally effective as measured by viewer recollection?

	Don't remember	*Remember vaguely*	*Remember key points*	*Total*
Commercial A	12	15	23	50
Commercial B	15	15	20	50
Total	27	30	43	100

Solution The null hypothesis is $H_0\colon p_{A1} = p_{B1}, p_{A2} = p_{B2}, p_{A3} = p_{B3}$, against the alternative that they are not all equal. The expected frequencies under the null hypothesis and the $(n_{ij} - e_{ij})^2/e_{i_j}$ terms, called contributions (*cont*) from cell (i,j) to the Q statistic, are shown below.

	Don't remember	*Remember vaguely*	*Remember key points*	*Total*
Commercial A	$X_{11} = 12$ $e_{11} = 13.5$ $cont = 0.17$	$X_{12} = 15$ $e_{12} = 15$ $cont = 0$	$X_{13} = 23$ $e_{13} = 21.5$ $cont = 0.10$	50
Commercial B	$X_{21} = 15$ $e_{21} = 13.5$ $cont = 0.17$	$X_{22} = 15$ $e_{22} = 15$ $cont = 0$	$X_{23} = 20$ $e_{23} = 21.5$ $cont = 0.10$	50
Total	27	30	43	100

The test statistic is $Q = 0.17 + 0 + 0.10 + 0.17 + 0 + 0.10 = 0.54$ and Table B of the Appendix with $df = 2$ shows $P > 0.50$. This implies that there is no significant difference between commercials A and B with respect to recollection by viewers. The STATXACT and MINITAB solutions are shown below. The answers agree with ours.

```
*************************************
  STATXACT SOLUTION TO EXAMPLE 6.1
*************************************
CHI-SQUARE TEST FOR INDEPENDENCE

Statistic based on the observed 2 by 3 table(x) :
   CH(X) : Pearson Chi-Square Statistic =  0.5426

Asymptotic p-value: (based on Chi-Square distribution with
2 df)
     Pr {CH(X) .GE.      0.5426 } = 0.7624

Exact p-value and point probability :
     Pr {CH(X) .GE.      0.5426 } = 0.7660
     Pr {CH(X) .EQ.      0.5426 } = 0.0513
```

```
*************************************
  MINITAB SOLUTION TO EXAMPLE 6.1
*************************************
```

Chi-Square Test: C1, C2, C3

Expected counts are printed below observed counts

	C1	C2	C3	Total
1	12	15	23	50
	13.50	15.00	21.50	
2	15	15	20	50
	13.50	15.00	21.50	
Total	27	30	43	100

```
Chi-Sq = 0.167 + 0.000 + 0.105 +
         0.167 + 0.000 + 0.105 = 0.543
DF = 2, P-Value = 0.762
```

ORDERED CATEGORIES

The three categories in Example 6.1 are actually ordered in terms of degree of recollection. In comparing two multinomial distributions when the categories are ordinal, we really are more interested in a one-sided alternative, specifically that the degree of recollection is greater for one commercial than the other, rather than the alternative that the degree of recollection is not the same for the two commercials. The chi-square test is appropriate only for the two-sided alternative. The Wilcoxon rank-sum test presented in Section 8.2 can be adapted to provide a test to compare two groups against a one-sided alternative. We explain this approach in the context of Eaxmple 6.2. This is very similar to what we did in Example 3.2 to calculate the Kendall tau coefficient.

Example 6.2 Two independent random samples of 10 business executives are taken, one sample from executives under 45 years of age, and the other from executives at least 45 years old. Each subject is then classified in terms of degree of risk aversion, Low, Medium or High, based on the results of a psychological test. For the data shown below, the research hypothesis of interest is that the younger business executives are more risk averse than their older counterparts.

	Degree of risk aversion			
Age	*Low*	*Medium*	*High*	*Total*
Under 45	2	3	5	10
Over 45	4	5	1	10
Total	6	8	6	20

Solution We call the Under 45 group the X sample and the Over 45 the Y sample. If we code (rank) the 3 degrees of risk aversion as $1 = \text{low}$, $2 = \text{medium}$, and $3 = \text{high}$, the six executives from the X and Y samples combined who were classified as Low (column one) are all tied at rank 1. If we use the midrank method to resolve the ties, each of these six executives (in the first column) would be assigned rank $(1+2+3+4+5+6)/6 = 21/6 = 3.5$. For the second column category (Medium), the midrank is $(7+8+\cdots+14)/8 = 84/8 = 10.5$, and for the third column, the midrank is $(15+16+\cdots+20)/6 = 105/6 = 17.5$. (Note that the midrank with integer ranks is always the average of the smallest and the largest ranks they would have had if they were not tied.) The value of the Wilcoxon rank-sum test statistic for the X sample is then $W_N = 2(3.5) + 3(10.5) + 5(17.5) = 126$. We can test the significance of this result using the normal approximation to the distribution of W_N for $m = 10$, $n = 10$, $N = 20$. The mean is $m(N+1)/2$ and the variance is calculated from (8.2.3) with the correction for ties. The reader can verify that the mean is 105 and the variance is 154.74, giving $Z = 1.688$ without a continuity correction and $Z = 1.648$ with a continuity correction. The upper tail P values from Table A are 0.046 and 0.050, respectively. This result does not lead to any firm conclusion at the 0.05 level.

Our first result agrees with the STATXACT output shown below. The output also shows the exact P value is 0.0785, which is not significant at the 0.05 level. This exact test is carried out using the conditional distribution of the cell counts given the column totals, which is a multiple hypergeometric distribution [see Lehmann (1975), p. 384]. The chi-square test for independence (a two-sided alternative) shows no significant difference between the two age groups.

```
                ********************************
                 STATXACT SOLUTION TO EXAMPLE 6.2
                ********************************

WILCOXON-MANN-WHITNEY TEST
[ 1 2 by 3 informative tables and sum of scores from row <row1  > ]

Summary of Exact distribution of WILCOXON-MANN-WHITNEY statistic:
  Min        Max       Mean       Std-dev     observed   Standardized
 63.00      147.0     105.0       12.44        126.0           1.688
 Mann-Whitney Statistic =    71.00

Asymptotic Inference:
 One-sided p-value: Pr {   Test Statistic .GE. Observed } =  0.0457
 Two-sided p-value: 2 * One-sided                         =  0.0914

Exact Inference:
 One-sided p-value: Pr {   Test Statistic .GE. Observed } =  0.0785
                    Pr {   Test Statistic .EQ. Observed } =  0.0563
 Two-sided p-value: Pr { | Test Statistic - Mean |
                         .GE.  | Observed - Mean |         =  0.1570
 Two-sided p-value: 2*One-Sided                           =  0.1570
```

The Wilcoxon rank-sum test can be considered a special case of a class of linear rank statistics of the form $T = \sum_j w_j X_{1j}$, where the w_j are some suitable scores or weights that are increasing in value. Different weights give rise to different linear rank statistics. For the Wilcoxon rank-sum test, the weights are the respective midranks.

Other possible weights could be based on the expected normal scores (Terry-Hoeffding) or inverse normal scores (van der Waerden). Graubard and Korn (1987) studied three classes of scores and made some recommendations. STATXACT 5.0 has options for calculating the power and sample size for any linear rank test, including the Wilcoxon rank-sum test.

PROBLEMS

14.1. An ongoing problem on college campuses is the instructor evaluation form. To aid in the interpretation of the results of such evaluations, a study was made to determine whether any relationship exists between the stage of a student's academic career and his attitude with respect to whether the academic work load in his courses was lighter than it should be, at the appropriate level, or heavier than it should be. A stratified random sample yielded the following results:

	Sophomore	*Junior*	*Senior*
Believe work load is lighter than it should be	5	8	11
Believe work load is at the appropriate level	30	35	40
Believe work load is heavier than it should be	25	17	9

(*a*) Test the null hypothesis that there is no association between the stage of a student's program and his attitude with respect to the appropriateness of the academic work load in his courses.

(*b*) Measure the degree of association.

14.2. A manufacturer produces units of a product in three 8-hour shifts: Day, Evening, and Night. Quality control teams check production lots for defects at the end of each shift by taking random samples. For the data below, do the three shifts have the same proportion of defects?

	Day	*Evening*	*Night*
Defects	70	60	80
Sample total	400	300	300

14.3. A group of 28 salespersons were rated on their sales presentations and then asked to view a training film on improving selling techniques. Each person was then rated a second time. For the data in Table 1 determine whether the training film has a positive effect on the ratings.

14.4. An employer wanted to find out if changing from his current health benefit policy to a prepaid policy would change hospitalization rates among his employees. A random sample of 100 employees was selected for the study. During the previous year under the current policy, 20 of them had been hospitalized and 80 had not been hospitalized. These

Table 1 Data for Problem 14.3

	Rating after film		
Rating before film	*Acceptable*	*Not acceptable*	*Total*
Acceptable	5	4	9
Nonacceptable	13	6	19
Total	18	10	28

same 100 employees were then placed on the prepaid policy and after one year, it was found that among the 20, 5 had been rehospitalized, and among the 80, 10 had been hospitalized. Test to see whether or not the prepaid policy reduces hospitalization rates among the employees.

14.5. A sample of five vaccinated and five unvaccinated cows were all exposed to a disease. Four cows contracted the disease, one from the vaccinated group and three from the nonvaccinated group. Determine whether the vaccination had a significant effect in protecting the cows against the disease.

14.6. A superintendent of schools is interested in revising the curriculum. He sends out questionnaires to 200 teachers: 100 respond *No* to the question "Do you think we should revise the curriculum?" The superintendent then held a weeklong workshop on curriculum improvement and sent out the same questionnaire to the same 200 teachers; this time 90 responded *No*. Eighty teachers responded *No* both times. Investigate whether the workshop significantly decreased the number of negative responses.

14.7. A retrospective study of death certificates was aimed at determining whether an association exists between a particular occupation and a certain neoplastic disease. In a certain geographical area over a period of time, some 1500 certificates listed the neoplastic disease as primary cause of death. For each of them, a matched control death certificate was selected, based on age, race, gender, county of residence, and date of death, and stating any cause of death other than the neoplastic disease. The occupation of each decedent was determined. Only one matched pair had both the case and control members in the specified occupation. There were 69 pairs in which the case pair member was in the specified occupation while the control member was not. There were 30 pairs in which the control member was in the occupation and the case pair member was not. In all of the remaining 1400 pairs, neither the case nor the control member was in the specified occupation. Test the null hypothesis that the proportion of case and control members in the occupation is the same.

Table 2 Data for Problem 14.8

	Firm asset size ($1000)			
	Less than 500	*500–2000*	*Over 2000*	*Total*
Debt less than equity	7	10	8	25
Debt greater than equity	10	18	9	37
Total	17	28	17	62

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

14.8. A financial consultant is interested in testing whether the proportion of debt that exceeds equity is the same irrespective of the magnitude of the firm's assets. Sixty-two firms are classified into three groups according to asset size and data in Table 2 are obtained on the numbers with debt greater then equity. Carry out the test.

14.9. In a study designed to investigate the relationship between age and degree of job satisfaction among clerical workers, a random sample of 100 clerical workers were interviewed and classified according to these characteristics as shown in the table below.

(*a*) Test whether there is any association between age and job satisfaction using the chi-square test.

(*b*) Calculate the contingency coefficient and the phi coefficient.

(*c*) Calculate Kendall's tau with correction for ties and test for association.

(*d*) Calculate the Goodman-Kruskal coefficient.

	Job satisfaction (1 = least satisfied)			
Age	*1*	*2*	*3*	*Total*
Under 25	8	7	5	20
25–39	12	8	20	40
40 and over	20	15	5	40
Total	40	30	30	100

14.10. A random sample of 135 U.S. citizens were asked their opinion about the current U.S. foreign policy in Afghanistan. Forty-three reported a negative opinion and the others were positive. These 135 persons were then put on a mailing list to receive an informative newsletter about U.S. foreign policy, and then asked their opinion a month later. At the time, 37 were opposed and 30 of these 37 originally had a positive opinion. Find the P value for the alternative that the probability of a change from negative to positive is greater than the corresponding probability of a change in the opposite direction.

14.11. A small random sample was used in an experiment to see how effective an informative newsletter was in persuading people to favor a flat income tax bill. Thirty persons were asked their opinion before receiving the letter and these same persons were then asked again after receiving the letter. Before the letter, 11 were in favor. Five were in favor both before and after receiving the newsletter, and 6 were opposed both times. Is there evidence that the letter is effective in persuading people to favor a flat tax?

14.12. Twenty married couples were selected at random from a large population and each person was asked privately whether the family would prefer to spend a week's summer vacation at the beach or at the mountains. The subjects were told to ignore factors such as relative cost and distance so that their preference would reflect only their expected pleasure from each type of vacation. The husband voted for the beach 7 times and his wife agreed 4 times. The husband voted for the mountains 13 times and his wife agreed 5 times. Determine whether family vacation preference is dominated by the husband.

14.13. A study was conducted to investigate whether high school experience with calculus has an effect on performance in first-year college calculus. A total of 686 students who had completed their first year of college calculus were classified according to their high school calculus experience as Zero (None), Brief (One semester), Year (Two semesters), and AP (Advanced Placement); these same students were then classified according to their grade in first year college calculus. Test to see whether high school calculus experience has an effect on college grade.

	High school calculus			
College grade	*Zero*	*Brief*	*Year*	*AP*
A	3	6	32	16
B	23	30	70	56
C	48	51	67	29
D	49	45	27	6
F	71	44	17	2

14.14. For the data in Problem 14.8, investigate whether firms with debt greater than equity tend to have more assets than other firms.

14.15. Derive the maximum likelihood estimators for the parameters in the likelihood function of (2.1)

14.16. Show that (2.2) is still the appropriate test statistic for independence in a two-way $r \times k$ contingency table when both the row and column totals are fixed.

14.17. Verify the equivalence of the expressions in (3.1), (3.2), and (3.3).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Appendix of Tables

SOURCE OF TABLES

The formats for Tables A, C through J, L through N, have been reprinted with permission from *Nonparametric Methods for Quantitative Analysis* (Third Edition) by Jean Dickinson Gibbons, copyright © 1976, 1985, 1997 by American Sciences Press, Inc., Syracuse, NY 13224-2144. Permission has also been obtained from the original sources, as shown for each table.

Table A Normal Distribution

Each table entry is the tail probability P, right tail from the value of z to plus infinity, and also left tail from minus infinity to $-z$, for all $P \leqslant .50$. Read down the first column to the first decimal value of z, and over to the correct column for the second decimal value; the number at the intersection is P.

z	*.00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*0.9*
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002

Source: Adapted from Table 1 of Pearson, E. S., and H. O. Hartley, eds. (1954), *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, Cambridge, England, with permission of the Biometrika Trustees.

Table B Chi-Square Distribution

Each table entry is the value of a chi-square random variable with v degrees of freedom such that its right-tail probability is the value given on the top row.

	Right-tail probability								
v	*0.95*	*0.90*	*0.50*	*0.25*	*0.10*	*0.05*	*0.01*	*0.005*	*0.001*
1	0.004	0.016	0.45	1.32	2.71	3.84	6.63	7.88	10.83
2	0.10	0.21	1.39	2.77	4.61	5.99	9.21	10.60	13.82
3	0.35	0.58	2.37	4.11	6.25	7.81	11.34	12.84	16.27
4	0.71	1.06	3.36	5.39	7.78	9.49	13.28	14.86	18.47
5	1.15	1.61	4.35	6.63	9.24	11.07	15.09	16.75	20.52
6	1.64	2.20	5.35	7.84	10.64	12.59	16.81	18.55	22.46
7	2.17	2.83	6.35	9.04	12.02	14.07	18.48	20.28	24.32
8	2.73	3.49	7.34	10.22	12.36	15.51	20.09	21.96	26.12
9	3.33	4.17	8.34	11.39	14.68	16.92	21.67	23.59	27.88
10	3.94	4.87	9.34	12.55	15.99	18.31	23.21	25.19	29.59
11	4.57	5.58	10.34	13.70	17.28	19.68	24.72	26.76	31.26
12	5.23	6.30	11.34	14.85	18.55	21.03	26.22	28.30	32.91
13	5.89	7.04	12.34	15.98	19.81	22.36	27.69	29.82	34.53
14	6.57	7.79	13.34	17.12	21.06	23.68	29.14	31.32	36.12
15	7.26	8.55	14.34	18.25	22.31	25.00	30.58	32.80	37.70
16	7.96	9.31	15.34	19.37	23.54	26.30	32.00	34.27	39.25
17	8.67	10.09	16.34	20.49	24.77	27.59	33.41	35.72	40.79
18	9.39	10.86	17.34	21.60	25.99	28.87	34.81	37.16	42.31
19	10.12	11.65	18.34	22.72	27.20	30.14	36.19	38.58	43.82
20	10.85	12.44	19.34	23.83	28.41	31.41	37.57	40.00	45.32
21	11.59	13.24	20.34	24.93	29.62	32.67	38.93	41.40	46.80
22	12.34	14.04	21.34	26.04	30.81	33.92	40.29	42.80	48.27
23	13.09	14.85	22.34	27.14	32.01	35.17	41.64	44.18	49.73
24	13.85	15.66	23.34	28.24	33.20	36.42	42.98	45.56	51.18
25	14.61	16.47	24.34	29.34	34.38	37.65	44.31	46.93	52.62
26	15.38	17.29	25.34	30.43	35.56	38.89	45.64	48.29	54.05
27	16.15	18.11	26.34	31.53	36.74	40.11	46.96	49.64	55.48
28	16.93	18.94	27.34	32.62	37.92	41.34	48.28	50.99	56.89
29	17.71	19.77	28.34	33.71	39.09	42.56	49.59	52.34	58.30
30	18.49	20.60	29.34	34.80	40.26	43.77	50.89	53.67	59.70

For $v > 30$, a right-tail or left-tail probability for Q a chi-square variable can be found from Table A with Z where $Z = \sqrt{2Q} - \sqrt{2v - 1}$.

Source: Adapted from Table 8 of Pearson, E. S. and H. O. Hartley, eds. (1954), *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, Cambridge, England, with permission of the Biometrika Trustees.

Table C Cumulative Binomial Distribution

Each table entry is the left-tail probability of x or less successes in n Bernoulli trials where θ is the probability of a success on each trial.

		θ								
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025
	1	.9975	.9900	.9775	.9600	.9375	.9100	.8775	.8400	.7975
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664
	1	.9928	.9720	.9392	.8960	.8438	.7840	.7182	.6480	.5748
	2	.9999	.9990	.9966	.9920	.9844	.9730	.9561	.9360	.9089
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915
	1	.9860	.9477	.8905	.8192	.7383	.6517	.5630	.4752	.3910
	2	.9995	.9963	.9880	.9728	.9492	.9163	.8735	.8208	.7585
	3	1.0000	.9999	.9995	.9984	.9961	.9919	.9850	.9744	.9590
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503
	1	.9774	.9185	.8352	.7373	.6328	.5282	.4284	.3370	.2562
	2	.9988	.9914	.9734	.9421	.8965	.8369	.7648	.6826	.5931
	3	1.0000	.9995	.9978	.9933	.9844	.9692	.9460	.9130	.8688
	4	1.0000	1.0000	.9999	.9997	.9990	.9976	.9947	.9898	.9815
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277
	1	.9672	.8857	.7765	.6554	.5339	.4202	.3191	.2333	.1636
	2	.9978	.9842	.9527	.9011	.8306	.7443	.6471	.5443	.4415
	3	.9999	.9987	.9941	.9830	.9624	.9295	.8826	.8208	.7447
	4	1.0000	.9999	.9996	.9984	.9954	.9891	.9777	.9590	.9308
	5	1.0000	1.0000	1.0000	.9999	.9998	.9993	.9982	.9959	.9917
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152
	1	.9556	.8503	.7166	.5767	.4449	.3294	.2338	.1586	.1024
	2	.9962	.9743	.9262	.8520	.7564	.6471	.5323	.4199	.3164
	3	.9998	.9973	.9879	.9667	.9294	.8740	.8002	.7102	.6083
	4	1.0000	.9998	.9988	.9953	.9871	.9712	.9444	.9037	.8471
	5	1.0000	1.0000	.9999	.9996	.9987	.9962	.9910	.9812	.9643
	6	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9994	.9984	.9963
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(Continued)

Table C (*Continued*)

		θ									
n	*x*	*.50*	*.55*	*.60*	*.65*	*.70*	*.75*	*.80*	*.85*	*.90*	*.95*
1	0	.5000	.4500	.4000	.3500	.3000	.2500	.2000	.1500	.1000	.0500
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0	.2500	.2025	.1600	.1225	.0900	.0625	.0400	.0225	.0100	.0025
	1	.7500	.6975	.6400	.5775	.5100	.4375	.3600	.2775	.1900	.0975
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	.1250	.0911	.0640	.0429	.0270	.0156	.0080	.0034	.0010	.0001
	1	.5000	.4252	.3520	.2818	.2160	.1562	.1040	.0608	.0280	.0072
	2	.8750	.8336	.7840	.7254	.6570	.5781	.4880	.3859	.2710	.1426
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	.0625	.0410	.0256	.0150	.0081	.0039	.0016	.0005	.0001	.0000
	1	.3125	.2415	.1792	.1265	.0837	.0508	.0272	.0120	.0037	.0005
	2	.6875	.6090	.5248	.4370	.3483	.2617	.1808	.1095	.0523	.0140
	3	.9375	.9085	.8704	.8215	.7599	.6836	.5904	.4780	.3439	.1855
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	.0312	.0185	.0102	.0053	.0024	.0010	.0003	.0001	.0000	.0000
	1	.1875	.1312	.0870	.0540	.0308	.0156	.0067	.0022	.0005	.0000
	2	.5000	.4069	.3174	.2352	.1631	.1035	.0579	.0266	.0086	.0012
	3	.8125	.7438	.6630	.5716	.4718	.3672	.2627	.1648	.0815	.0226
	4	.9688	.9497	.9222	.8840	.8319	.7627	.6723	.5563	.4095	.2262
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	.0156	.0083	.0041	.0018	.0007	.0002	.0001	.0000	.0000	.0000
	1	.1094	.0692	.0410	.0223	.0109	.0046	.0016	.0004	.0001	.0000
	2	.3438	.2553	.1792	.1174	.0705	.0376	.0170	.0059	.0013	.0001
	3	.6562	.5585	.4557	.3529	.2557	.1694	.0989	.0473	.0158	.0022
	4	.8906	.8364	.7667	.6809	.5798	.4661	.3446	.2235	.1143	.0328
	5	.9844	.9723	.9533	.9246	.8824	.8220	.7379	.6229	.4686	.2649
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	.0078	.0037	.0016	.0006	.0002	.0001	.0000	.0000	.0000	.0000
	1	.0625	.0357	.0188	.0090	.0038	.0013	.0004	.0001	.0000	.0000
	2	.2266	.1529	.0963	.0556	.0288	.0129	.0047	.0012	.0002	.0000
	3	.5000	.3917	.2898	.1998	.1260	.0706	.0333	.0121	.0027	.0002
	4	.7734	.6836	.5801	.4677	.3529	.2436	.1480	.0738	.0257	.0038
	5	.9375	.8976	.8414	.7662	.6706	.5551	.4233	.2834	.1497	.0444
	6	.9922	.9848	.9720	.9510	.9176	.8665	.7903	.6794	.5217	.3017
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ								
n	*x*	.05	.10	.15	.20	.25	.30	.35	.40	.45
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084
	1	.9428	.8131	.6572	.5033	.3671	.2553	.1691	.1064	.0632
	2	.9942	.9619	.8948	.7969	.6785	.5518	.4278	.3154	.2201
	3	.9996	.9950	.9786	.9437	.8862	.8059	.7064	.5941	.4770
	4	1.0000	.9996	.9971	.9896	.9727	.9420	.8939	.8263	.7396
	5	1.0000	1.0000	.9998	.9988	.9958	.9887	.9747	.9502	.9115
	6	1.0000	1.0000	1.0000	.9999	.9996	.9987	.9964	.9915	.9819
	7	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9993	.9983
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046
	1	.9288	.7748	.5995	.4362	.3003	.1960	.1211	.0705	.0385
	2	.9916	.9470	.8591	.7382	.6007	.4628	.3373	.2318	.1495
	3	.9994	.9917	.9661	.9144	.8343	.7297	.6089	.4826	.3614
	4	1.0000	.9991	.9944	.9804	.9511	.9012	.8283	.7334	.6214
	5	1.0000	.9999	.9994	.9969	.9900	.9747	.9464	.9006	.8342
	6	1.0000	1.0000	1.0000	.9997	.9987	.9957	.9888	.9750	.9502
	7	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9986	.9962	.9909
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9992
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025
	1	.9139	.7361	.5443	.3758	.2440	.1493	.0860	.0464	.0233
	2	.9885	.9298	.8202	.6778	.5256	.3828	.2616	.1673	.0996
	3	.9990	.9872	.9500	.8791	.7759	.6496	.5138	.3823	.2660
	4	.9999	.9984	.9901	.9672	.9219	.8497	.7515	.6331	.5044
	5	1.0000	.9999	.9986	.9936	.9803	.9527	.9051	.8338	.7384
	6	1.0000	1.0000	.9999	.9991	.9965	.9894	.9740	.9452	.8980
	7	1.0000	1.0000	1.0000	.9999	.9996	.9984	.9952	.9877	.9726
	8	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9995	.9983	.9955
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014
	1	.8981	.6974	.4922	.3221	.1971	.1130	.0606	.0302	.0139
	2	.9848	.9104	.7788	.6174	.4552	.3127	.2001	.1189	.0652
	3	.9984	.9815	.9306	.8389	.7133	.5696	.4256	.2963	.1911
	4	.9999	.9972	.9841	.9496	.8854	.7897	.6683	.5328	.3971
	5	1.0000	.9997	.9973	.9883	.9657	.9218	.8513	.7535	.6331
	6	1.0000	1.0000	.9997	.9980	.9924	.9784	.9499	.9006	.8262
	7	1.0000	1.0000	1.0000	.9998	.9988	.9957	.9878	.9707	.9390
	8	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9980	.9941	.9852
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9993	.9978
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ									
n	*x*	*.50*	*.55*	*.60*	*.65*	*.70*	*.75*	*.80*	*.85*	*.90*	*.95*
8	0	.0039	.0017	.0007	.0002	.0001	.0000	.0000	.0000	.0000	.0000
	1	.0352	.0181	.0085	.0036	.0013	.0004	.0001	.0000	.0000	.0000
	2	.1445	.0885	.0498	.0253	.0113	.0042	.0012	.0002	.0000	.0000
	3	.3633	.2604	.1737	.1061	.0580	.0273	.0104	.0029	.0004	.0000
	4	.6367	.5230	.4059	.2936	.1941	.1138	.0563	.0214	.0050	.0004
	5	.8555	.7799	.6846	.5722	.4482	.3215	.2031	.1052	.0381	.0058
	6	.9648	.9368	.8936	.8309	.7447	.6329	.4967	.3428	.1869	.0572
	7	.9961	.9916	.9832	.9681	.9424	.8999	.8322	.7275	.5695	.3366
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	.0020	.0008	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0195	.0091	.0038	.0014	.0004	.0001	.0000	.0000	.0000	.0000
	2	.0898	.0498	.0250	.0112	.0043	.0013	.0003	.0000	.0000	.0000
	3	.2539	.1658	.0994	.0536	.0253	.0100	.0031	.0006	.0001	.0000
	4	.5000	.3786	.2666	.1717	.0988	.0489	.0196	.0056	.0009	.0000
	5	.7461	.6386	.5174	.3911	.2703	.1657	.0856	.0339	.0083	.0006
	6	.9102	.8505	.7682	.6627	.5372	.3993	.2618	.1409	.0530	.0084
	7	.9805	.9615	.9295	.8789	.8040	.6997	.5638	.4005	.2252	.0712
	8	.9980	.9954	.9899	.9793	.9596	.9249	.8658	.7684	.6126	.3698
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	.0010	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0107	.0045	.0017	.0005	.0001	.0000	.0000	.0000	.0000	.0000
	2	.0547	.0274	.0123	.0048	.0016	.0004	.0001	.0000	.0000	.0000
	3	.1719	.1020	.0548	.0260	.0106	.0035	.0009	.0001	.0000	.0000
	4	.3770	.2616	.1662	.0949	.0473	.0197	.0064	.0014	.0001	.0000
	5	.6230	.4956	.3669	.2485	.1503	.0781	.0328	.0099	.0016	.0001
	6	.8281	.7340	.6177	.4862	.3504	.2241	.1209	.0500	.0128	.0010
	7	.9453	.9004	.8327	.7384	.6172	.4744	.3222	.1798	.0702	.0115
	8	.9893	.9767	.9536	.9140	.8507	.7560	.6242	.4557	.2639	.0861
	9	.9990	.9975	.9940	.9865	.9718	.9437	.8926	.8031	.6513	.4013
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0	.0005	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0059	.0022	.0007	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0327	.0148	.0059	.0020	.0006	.0001	.0000	.0000	.0000	.0000
	3	.1133	.0610	.0293	.0122	.0043	.0012	.0002	.0000	.0000	.0000
	4	.2744	.1738	.0994	.0501	.0216	.0076	.0020	.0003	.0000	.0000
	5	.5000	.3669	.2465	.1487	.0782	.0343	.0117	.0027	.0003	.0000
	6	.7256	.6029	.4672	.3317	.2103	.1146	.0504	.0159	.0028	.0001
	7	.8867	.8089	.7037	.5744	.4304	.2867	.1611	.0694	.0185	.0016
	8	.9673	.9348	.8811	.7999	.6873	.5448	.3826	.2212	.0896	.0152
	9	.9941	.9861	.9698	.9394	.8870	.8029	.6779	.5078	.3026	.1019
	10	.9995	.9986	.9964	.9912	.9802	.9578	.9141	.8327	.6862	.4312
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ								
n	*x*	.05	.10	.15	.20	.25	.30	.35	.40	.45
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008
	1	.8816	.6590	.4435	.2749	.1584	.0850	.0424	.0196	.0083
	2	.9804	.8891	.7358	.5583	.3907	.2528	.1513	.0834	.0421
	3	.9978	.9744	.9078	.7946	.6488	.4925	.3467	.2253	.1345
	4	.9998	.9957	.9761	.9274	.8424	.7237	.5833	.4382	.3044
	5	1.0000	.9995	.9954	.9806	.9456	.8822	.7873	.6652	.5269
	6	1.0000	.9999	.9993	.9961	.9857	.9614	.9154	.8418	.7393
	7	1.0000	1.0000	.9999	.9994	.9972	.9905	.9745	.9427	.8883
	8	1.0000	1.0000	1.0000	.9999	.9996	.9983	.9944	.9847	.9644
	9	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9992	.9972	.9921
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9989
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004
	1	.8646	.6213	.3983	.2336	.1267	.0637	.0296	.0126	.0049
	2	.9755	.8661	.7296	.5017	.3326	.2025	.1132	.0579	.0269
	3	.9969	.9658	.9033	.7473	.5843	.4206	.2783	.1686	.0929
	4	.9997	.9935	.9740	.9009	.7940	.6543	.5005	.3530	.2279
	5	1.0000	.9991	.9947	.9700	.9198	.8346	.7159	.5744	.4268
	6	1.0000	.9999	.9987	.9930	.9757	.9376	.8705	.7712	.6437
	7	1.0000	1.0000	.9998	.9988	.9944	.9818	.9538	.9023	.8212
	8	1.0000	1.0000	1.0000	.9998	.9990	.9960	.9874	.9679	.9302
	9	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9975	.9922	.9797
	10	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9987	.9959
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9995
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002
	1	.8470	.5846	.3567	.1979	.1010	.0475	.0205	.0081	.0029
	2	.9699	.8416	.6479	.4481	.2811	.1608	.0839	.0398	.0170
	3	.9958	.9559	.8535	.6982	.5213	.3552	.2205	.1243	.0632
	4	.9996	.9908	.9533	.8702	.7415	.5842	.4227	.2793	.1672
	5	1.0000	.9985	.9885	.9561	.8883	.7805	.6405	.4859	.3373
	6	1.0000	.9998	.9978	.9884	.9617	.9067	.8164	.6925	.5461
	7	1.0000	1.0000	.9997	.9976	.9897	.9685	.9247	.8499	.7414
	8	1.0000	1.0000	1.0000	.9996	.9978	.9917	.9757	.9417	.8811
	9	1.0000	1.0000	1.0000	1.0000	.9997	.9983	.9940	.9825	.9574
	10	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9989	.9961	.9886
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9978
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ									
n	*x*	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
12	0	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0032	.0011	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0193	.0079	.0028	.0008	.0002	.0000	.0000	.0000	.0000	.0000
	3	.0730	.0356	.0153	.0056	.0017	.0004	.0001	.0000	.0000	.0000
	4	.1938	.1117	.0573	.0255	.0095	.0028	.0006	.0001	.0000	.0000
	5	.3872	.2607	.1582	.0846	.0386	.0143	.0039	.0007	.0001	.0000
	6	.6128	.4731	.3348	.2127	.1178	.0544	.0194	.0046	.0005	.0000
	7	.8062	.6956	.5618	.4167	.2763	.1576	.0726	.0239	.0043	.0002
	8	.9270	.8655	.7747	.6533	.5075	.3512	.2054	.0922	.0256	.0022
	9	.9807	.9579	.9166	.8487	.7472	.6093	.4417	.2642	.1109	.0196
	10	.9968	.9917	.9804	.9576	.9150	.8416	.7251	.5565	.3410	.1184
	11	.9998	.9992	.9978	.9943	.9862	.9683	.9313	.8578	.7176	.4596
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	0	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0017	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0112	.0041	.0013	.0003	.0001	.0000	.0000	.0000	.0000	.0000
	3	.0461	.0203	.0078	.0025	.0007	.0001	.0000	.0000	.0000	.0000
	4	.1334	.0698	.0321	.0126	.0040	.0010	.0002	.0000	.0000	.0000
	5	.2905	.1788	.0977	.0462	.0182	.0056	.0012	.0002	.0000	.0000
	6	.5000	.3563	.2288	.1295	.0624	.0243	.0070	.0013	.0001	.0000
	7	.7095	.5732	.4256	.2841	.1654	.0802	.0300	.0053	.0009	.0000
	8	.8666	.7721	.6470	.4995	.3457	.2060	.0991	.0260	.0065	.0003
	9	.9539	.9071	.8314	.7217	.5794	.4157	.2527	.0967	.0342	.0031
	10	.9888	.9731	.9421	.8868	.7975	.6674	.4983	.2704	.1339	.0245
	11	.9983	.9951	.9874	.9704	.9363	.8733	.7664	.6017	.3787	.1354
	12	.9999	.9996	.9987	.9963	.9903	.9762	.9450	.8791	.7458	.4867
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0009	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0065	.0022	.0006	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0287	.0114	.0039	.0011	.0002	.0000	.0000	.0000	.0000	.0000
	4	.0898	.0462	.0175	.0060	.0017	.0003	.0000	.0000	.0000	.0000
	5	.2120	.1189	.0583	.0243	.0083	.0022	.0004	.0000	.0000	.0000
	6	.3953	.2586	.1501	.0753	.0315	.0103	.0024	.0003	.0000	.0000
	7	.6047	.4539	.3075	.1836	.0933	.0383	.0116	.0022	.0002	.0000
	8	.7880	.6627	.5141	.3595	.2195	.1117	.0439	.0115	.0015	.0000
	9	.9102	.8328	.7207	.5773	.4158	.2585	.1298	.0467	.0092	.0004
	10	.9713	.9368	.8757	.7795	.6448	.4787	.3018	.1465	.0441	.0042
	11	.9935	.9830	.9602	.9161	.8392	.7189	.5519	.3521	.1584	.0301
	12	.9991	.9971	.9919	.9795	.9525	.8990	.8021	.6433	.4154	.1530
	13	.9999	.9998	.9992	.9976	.9932	.9822	.9560	.8972	.7712	.5123
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ								
n	*x*	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001
	1	.8290	.5490	.3186	.1671	.0802	.0353	.0142	.0052	.0017
	2	.9638	.8159	.6042	.3980	.2361	.1268	.0617	.0271	.0107
	3	.9945	.9444	.8227	.6482	.4613	.2969	.1727	.0905	.0424
	4	.9994	.9873	.9383	.8358	.6865	.5155	.3519	.2173	.1204
	5	.9999	.9978	.9832	.9389	.8516	.7216	.5643	.4032	.2608
	6	1.0000	.9997	.9964	.9819	.9434	.8689	.7548	.6098	.4522
	7	1.0000	1.0000	.9994	.9958	.9927	.9500	.8868	.7869	.6535
	8	1.0000	1.0000	.9999	.9992	.9958	.9848	.9578	.9050	.8182
	9	1.0000	1.0000	1.0000	.9999	.9992	.9963	.9876	.9662	.9231
	10	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9972	.9907	.9745
	11	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9995	.9981	.9937
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9989
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16	0	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001
	1	.8108	.5147	.2839	.1407	.0635	.0261	.0098	.0033	.0010
	2	.9571	.7892	.5614	.3518	.1971	.0994	.0451	.0183	.0066
	3	.9930	.9316	.7899	.5981	.4050	.2459	.1339	.0651	.0281
	4	.9991	.9830	.9209	.7982	.6302	.4499	.2892	.1666	.0853
	5	.9999	.9967	.9765	.9183	.8103	.6598	.4900	.3288	.1976
	6	1.0000	.9995	.9944	.9733	.9204	.8247	.6881	.5272	.3660
	7	1.0000	.9999	.9989	.9930	.9729	.9256	.8406	.7161	.5629
	8	1.0000	1.0000	.9998	.9985	.9925	.9743	.9329	.8577	.7441
	9	1.0000	1.0000	1.0000	.9998	.9984	.9929	.9771	.9417	.8759
	10	1.0000	1.0000	1.0000	1.0000	.9997	.9984	.9938	.9809	.9514
	11	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9987	.9951	.9851
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9991	.9965
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9991	.9994
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

n	x	θ = .50	.55	.60	.65	.70	.75	.80	.85	.90	.95
15	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0037	.0011	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0176	.0063	.0019	.0005	.0001	.0000	.0000	.0000	.0000	.0000
	4	.0592	.0255	.0093	.0028	.0007	.0001	.0000	.0000	.0000	.0000
	5	.1509	.0769	.0338	.0124	.0037	.0008	.0001	.0000	.0000	.0000
	6	.3036	.1818	.0950	.0422	.0152	.0042	.0008	.0001	.0000	.0000
	7	.5000	.3465	.2131	.1132	.0500	.0173	.0042	.0006	.0000	.0000
	8	.6964	.5478	.3902	.2452	.1311	.0566	.0181	.0036	.0003	.0000
	9	.8491	.7392	.5968	.4357	.2784	.1484	.0611	.0168	.0022	.0001
	10	.9408	.8796	.7827	.6481	.4845	.3135	.1642	.0617	.0127	.0006
	11	.9824	.9576	.9095	.8273	.7031	.5387	.3518	.1773	.0556	.0055
	12	.9963	.9893	.9729	.9383	.8732	.7639	.6020	.3958	.1841	.0362
	13	.9995	.9983	.9948	.9858	.9647	.9198	.8329	.6814	.4510	.1710
	14	1.0000	.9999	.9995	.9984	.9953	.9866	.9648	.9126	.7941	.5367
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.000	1.0000	1.0000	1.0000
16	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0021	.0006	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0106	.0035	.0009	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	4	.0384	.0149	.0049	.0013	.0003	.0000	.0000	.0000	.0000	.0000
	5	.1051	.0486	.0191	.0062	.0016	.0003	.0000	.0000	.0000	.0000
	6	.2272	.1241	.0583	.0229	.0071	.0016	.0002	.0000	.0000	.0000
	7	.4018	.2559	.1423	.0671	.0257	.0075	.0015	.0002	.0000	.0000
	8	.5982	.4371	.2839	.1594	.0744	.0271	.0070	.0011	.0001	.0000
	9	.7728	.6340	.4728	.3119	.1753	.0796	.0267	.0056	.0005	.0000
	10	.8949	.8024	.6712	.5100	.3402	.1897	.0817	.0235	.0033	.0001
	11	.9616	.9147	.8334	.7108	.5501	.3698	.2018	.0791	.0170	.0009
	12	.9894	.9719	.9349	.8661	.7541	.5950	.4019	.2101	.0684	.0070
	13	.9979	.9934	.9817	.9549	.9006	.8729	.6482	.4386	.2108	.0429
	14	.9997	.9990	.9967	.9902	.9739	.9365	.8593	.7161	.4853	.1892
	15	1.0000	.9999	.9997	.9990	.9967	.9900	.9719	.9257	.8147	.5599
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ								
n	*x*	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.45*
17	0	.4181	.1668	.0631	.0225	.0075	.0023	.0007	.0002	.0000
	1	.7922	.4818	.2525	.1182	.0501	.0193	.0067	.0021	.0006
	2	.9497	.7618	.5198	.3096	.1637	.0774	.0327	.0123	.0041
	3	.9912	.9174	.7556	.5489	.3530	.2019	.1028	.0464	.0184
	4	.9988	.9779	.9013	.7582	.5739	.3887	.2348	.1260	.0596
	5	.9999	.9953	.9681	.8943	.7653	.5968	.4197	.2639	.1471
	6	1.0000	.9992	.9917	.9623	.8929	.7752	.6188	.4478	.2902
	7	1.0000	.9999	.9983	.9891	.9598	.8954	.7872	.6405	.4743
	8	1.0000	1.0000	.9997	.9974	.9876	.9597	.9006	.8011	.6626
	9	1.0000	1.0000	1.0000	.9995	.9969	.9873	.9617	.9081	.8166
	10	1.0000	1.0000	1.0000	.9999	.9994	.9968	.9880	.9652	.9174
	11	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9970	.9894	.9699
	12	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9975	.9914
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9995	.9981
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	0	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000
	1	.7735	.4503	.2241	.0991	.0395	.0142	.0046	.0013	.0003
	2	.9419	.7338	.4797	.2713	.1353	.0600	.0236	.0082	.0025
	3	.9891	.9018	.7202	.5010	.3057	.1646	.0783	.0328	.0120
	4	.9985	.9718	.8794	.7164	.5187	.3327	.1886	.0942	.0411
	5	.9998	.9936	.9581	.8671	.7175	.5344	.3550	.2088	.1077
	6	1.0000	.9988	.9882	.9487	.8610	.7217	.5491	.3743	.2258
	7	1.0000	.9998	.9973	.9837	.9431	.8593	.7283	.5634	.3915
	8	1.0000	1.0000	.9995	.9957	.9807	.9404	.8609	.7368	.5778
	9	1.0000	1.0000	.9999	.9991	.9946	.9790	.9403	.8653	.7473
	10	1.0000	1.0000	1.0000	.9998	.9988	.9939	.9788	.9424	.8720
	11	1.0000	1.0000	1.0000	1.0000	.9998	.9986	.9938	.9797	.9463
	12	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9986	.9942	.9817
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9987	.9951
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9990
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ									
n	*x*	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
17	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0012	.0003	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0064	.0019	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	4	.0245	.0086	.0025	.0006	.0001	.0000	.0000	.0000	.0000	.0000
	5	.0717	.0301	.0106	.0030	.0007	.0001	.0000	.0000	.0000	.0000
	6	.1662	.0826	.0348	.0120	.0032	.0006	.0001	.0000	.0000	.0000
	7	.3145	.1834	.0919	.0383	.0127	.0031	.0005	.0000	.0000	.0000
	8	.5000	.3374	.1989	.0994	.0403	.0124	.0026	.0003	.0000	.0000
	9	.6855	.5257	.3595	.2128	.1046	.0402	.0109	.0017	.0001	.0000
	10	.8338	.7098	.5522	.3812	.2248	.1071	.0377	.0083	.0008	.0000
	11	.9283	.8529	.7361	.5803	.4032	.2347	.1057	.0319	.047	.0001
	12	.9755	.9404	.8740	.7652	.6113	.4261	.2418	.0987	.0221	.0012
	13	.9936	.9816	.9536	.8972	.7981	.6470	.4511	.2444	.0826	.0088
	14	.9988	.9959	.9877	.9673	.9226	.8363	.6904	.4802	.2382	.0503
	15	.9999	.9994	.9979	.9933	.9807	.9499	.8818	.7475	.5182	.2078
	16	1.0000	1.0000	.9998	.9993	.9977	.9925	.9775	.9369	.8332	.5819
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0007	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0038	.0010	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	4	.0154	.0049	.0013	.0003	.0000	.0000	.0000	.0000	.0000	.0000
	5	.0481	.0183	.0058	.0014	.0003	.0000	.0000	.0000	.0000	.0000
	6	.1189	.0537	.0203	.0062	.0014	.0002	.0000	.0000	.0000	.0000
	7	.2403	.1280	.0576	.0212	.0061	.0012	.0002	.0000	.0000	.0000
	8	.4073	.2527	.1347	.0597	.0210	.0054	.0009	.0001	.0000	.0000
	9	.5927	.4222	.2632	.1391	.0596	.0193	.0043	.0005	.0000	.0000
	10	.7597	.6085	.4366	.2717	.1407	.0569	.0163	.0027	.0002	.0000
	11	.8811	.7742	.6257	.4509	.2783	.1390	.0513	.0118	.0012	.0000
	12	.9519	.8923	.7912	.6450	.4656	.2825	.1329	.0419	.0064	.0002
	13	.9846	.9589	.9058	.8114	.6673	.4813	.2836	.1206	.0282	.0015
	14	.9962	.9880	.9672	.9217	.8354	.6943	.4990	.2798	.0982	.0109
	15	.9993	.9975	.9918	.9764	.9400	.8647	.7287	.5203	.2662	.0581
	16	.9999	.9997	.9987	.9954	.9858	.9605	.9009	.7759	.5497	.2265
	17	1.0000	1.0000	.9999	.9996	.9984	.9944	.9820	.9464	.8499	.6028
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ								
n	*x*	.05	.10	.15	.20	.25	.30	.35	.40	.45
19	0	.3774	.1351	.0456	.0144	.0042	.0011	.0003	.0001	.0000
	1	.7547	.4203	.1985	.0829	.0310	.0104	.0031	.0008	.0002
	2	.9335	.7054	.4413	.2369	.1113	.0462	.0170	.0055	.0015
	3	.9868	.8850	.6841	.4551	.2631	.1332	.0591	.0230	.0077
	4	.9980	.9648	.8556	.6733	.4654	.2822	.1500	.0696	.0280
	5	.9998	.9914	.9463	.8369	.6678	.4739	.2968	.1629	.0777
	6	1.0000	.9983	.9837	.9324	.8251	.6655	.4812	.3081	.1727
	7	1.0000	.9997	.9959	.9767	.9225	.8180	.6656	.4878	.3169
	8	1.0000	1.0000	.9992	.9933	.9713	.9161	.8145	.6675	.4940
	9	1.0000	1.0000	.9999	.9984	.9911	.9674	.9125	.8139	.6710
	10	1.0000	1.0000	1.0000	.9997	.9977	.9895	.9653	.9115	.8159
	11	1.0000	1.0000	1.0000	1.0000	.9995	.9972	.9886	.9648	.9129
	12	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9969	.9884	.9658
	13	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9969	.9891
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9972
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9995
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000
	1	.7358	.3917	.1756	.0692	.0243	.0076	.0021	.0005	.0001
	2	.9245	.6769	.4049	.2061	.0913	.0355	.0121	.0036	.0009
	3	.9841	.8670	.6477	.4114	.2252	.1071	.0444	.0160	.0049
	4	.9974	.9568	.8298	.6296	.4148	.2375	.1182	.0510	.0189
	5	.9997	.9887	.9327	.8042	.6172	.4164	.2454	.1256	.0553
	6	1.0000	.9976	.9781	.9133	.7858	.6080	.4166	.2500	.1299
	7	1.0000	.9996	.9941	.9679	.8982	.7723	.6010	.4159	.2520
	8	1.0000	.9999	.9987	.9900	.9591	.8867	.7624	.5956	.4143
	9	1.0000	1.0000	.9998	.9974	.9861	.9520	.8782	.7553	.5914
	10	1.0000	1.0000	1.0000	.9994	.9961	.9829	.9468	.8725	.7507
	11	1.0000	1.0000	1.0000	.9999	.9991	.9949	.9804	.9435	.8692
	12	1.0000	1.0000	1.0000	1.0000	.9998	.9987	.9940	.9790	.9420
	13	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9985	.9935	.9786
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9984	.9936
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9985
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(*Continued*)

Table C (*Continued*)

		θ									
n	*x*	*.50*	*.55*	*.60*	*.65*	*.70*	*.75*	*.80*	*.85*	*.90*	*95*
19	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0004	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0022	.0005	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	4	.0096	.0028	.0006	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	5	.0318	.0109	.0031	.0007	.0001	.0000	.0000	.0000	.0000	.0000
	6	.0835	.0342	.0116	.0031	.0006	.0001	.0000	.0000	.0000	.0000
	7	.1796	.0871	.0352	.0114	.0028	.0005	.0000	.0000	.0000	.0000
	8	.3238	.1841	.0885	.0347	.0105	.0023	.0003	.0000	.0000	.0000
	9	.5000	.3290	.1861	.0875	.0326	.0089	.0016	.0001	.0000	.0000
	10	.6762	.5060	.3325	.1855	.0839	.0287	.0067	.0008	.0000	.0000
	11	.8204	.6831	.5122	.3344	.1820	.0775	.0233	.0041	.0003	.0000
	12	.9165	.8273	.6919	.5188	.3345	.1749	.0676	.0163	.0017	.0000
	13	.9682	.9223	.8371	.7032	.5261	.3322	.1631	.0537	.0086	.0002
	14	.9904	.9720	.9304	.8500	.7178	.5346	.3267	.1444	.0352	.0020
	15	.9978	.9923	.9770	.9409	.8668	.7369	.5449	.3159	.1150	.0132
	16	.9996	.9985	.9945	.9830	.9538	.8887	.7631	.5587	.2946	.0665
	17	1.0000	.9998	.9992	.9969	.9896	.9690	.9171	.8015	.5797	.2453
	18	1.0000	1.0000	.9999	.9997	.9989	.9958	.9856	.9544	.8649	.6226
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0013	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	4	.0059	.0015	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	5	.0207	.0064	.0016	.0003	.0000	.0000	.0000	.0000	.0000	.0000
	6	.0577	.0214	.0065	.0015	.0003	.0000	.0000	.0000	.0000	.0000
	7	.1316	.0580	.0210	.0060	.0013	.0002	.0000	.0000	.0000	.0000
	8	.2517	.1308	.0565	.0196	.0051	.0009	.0001	.0000	.0000	.0000
	9	.4119	.2493	.1275	.0532	.0171	.0039	.0006	.0000	.0000	.0000
	10	.5881	.4086	.2447	.1218	.0480	.0139	.0026	.0002	.0000	.0000
	11	.7483	.5857	.4044	.2376	.1133	.0409	.0100	.0013	.0001	.0000
	12	.8684	.7480	.5841	.3990	.2277	.1018	.0321	.0059	.0004	.0000
	13	.9423	.8701	.7500	.5834	.3920	.2142	.0867	.0219	.0024	.0000
	14	.9793	.9447	.8744	.7546	.5836	.3828	.1958	.0673	.0113	.0003
	15	.9941	.9811	.9490	.8818	.7625	.5852	.3704	.1702	.0432	.0026
	16	.9987	.9951	.9840	.9556	.8929	.7748	.5886	.3523	.1330	.0159
	17	.9998	.9991	.9964	.9879	.9645	.9087	.7939	.5951	.3231	.0755
	18	1.0000	.9999	.9995	.9979	.9924	.9757	.9308	.8244	.6083	.2642
	19	1.0000	1.0000	1.0000	.9998	.9992	.9968	.9885	.9612	.8784	.6415
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Source: Adapted from Table 2 of *Tables of the Binomial Distribution,* (*January 1950 with Corrigenda* 1952 and 1958), National Bureau of Standards, U.S. Government Printing Office, Washington, D.C., with permission.

Table D Total Number of Runs Distribution
Each table entry labeled P is the tail probability from each extreme to the value of R, the total number of runs in a sequence of $n = n_1 + n_2$ symbols of two types for $n_1 \leqslant n_2$.

Left-tail probabilities

n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P
2	2	2	.333	2	18	2	.011	3	14	2	.003	4	10	2	.002
2	3	2	.200			3	.105			3	.025			3	.014
		3	.500			4	.284			4	.101			4	.068
2	4	2	.133	3	3	2	.100			5	.350			5	.203
		3	.400			3	.300	3	15	2	.002			6	.419
2	5	2	.095	3	4	2	.057			3	.022	4	11	2	.001
		3	.333			3	.200			4	.091			3	.011
2	6	2	.071	3	5	2	.036			5	.331			4	.055
		3	.286			3	.143	3	16	2	.002			5	.176
2	7	2	.056			4	.429			3	.020			6	.374
		3	.250	3	6	2	.024			4	.082	4	12	2	.001
2	8	2	.044			3	.107			5	.314			3	.009
		3	.222			4	.345	3	17	2	.002			4	.045
2	9	2	.036	3	7	2	.017			3	.018			5	.154
		3	.200			3	.083			4	.074			6	.335
		4	.491			4	.283			5	.298	4	13	2	.001
2	10	2	.030	3	8	2	.012	4	4	2	.029			3	.007
		3	.182			3	.067			3	.114			4	.037
		4	.455			4	.236			4	.371			5	.136
2	11	2	.026	3	9	2	.009	4	5	2	.016			6	.302
		3	.167			3	.055			3	.071	4	14	2	.001
		4	.423			4	.200			4	.262			3	.006
2	12	2	.022			5	.491			5	.500			4	.031
		3	.154	3	10	2	.007	4	6	2	.010			5	.121
		4	.396			3	.045			3	.048			6	.274
2	13	2	.019			4	.171			4	.190	4	15	2	.001
		3	.143			5	.455			5	.405			3	.005
		4	.371	3	11	2	.005	4	7	2	.006			4	.027
2	14	2	.017			3	.038			3	.033			5	.108
		3	.133			4	.148			4	.142			6	.249
		4	.350			5	.423			5	.333	4	16	2	.000
2	15	2	.015	3	12	2	.004	4	8	2	.004			3	.004
		3	.125			3	.033			3	.024			4	.023
		4	.331			4	.130			4	.109			5	.097
2	16	2	.013			5	.396			5	.279			6	.227
		3	.118	3	13	2	.004	4	9	2	.003	5	5	2	.008
		4	.314			3	.029			3	.018			3	.040
2	17	2	.012			4	.114			4	.085			4	.167
		3	.111			5	.371			5	.236			5	.357
		4	.298							6	.471				

(*Continued*)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table D (*Continued*)

Left-tail probabilities															
n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P
5	6	2	.004	5	14	2	.000	6	11	2	.000	7	9	2	.000
		3	.024			3	.002			3	.001			3	.001
		4	.110			4	.011			4	.009			4	.010
		5	.262			5	.044			5	.036			5	.035
5	7	2	.003			6	.125			6	.108			6	.108
		3	.015			7	.299			7	.242			7	.231
		4	.076			8	.496			8	.436			8	.427
		5	.197	5	15	2	.000	6	12	2	.000	7	10	2	.000
		6	.424			3	.001			3	.001			3	.001
5	8	2	.002			4	.009			4	.007			4	.006
		3	.010			5	.037			5	.028			2	.024
		4	.054			6	.108			6	.087			6	.080
		5	.152			7	.272			7	.205			7	.182
		6	.347			8	.460			8	.383			8	.355
5	9	2	.001	6	6	2	.002	6	13	2	.000	7	11	2	.000
		3	.007			3	.013			3	.001			3	.001
		4	.039			4	.067			4	.005			4	.004
		5	.119			5	.175			5	.022			5	.018
		6	.287			6	.392			6	.070			6	.060
5	10	2	.001	6	7	2	.001			7	.176			7	.145
		3	.005			3	.008			8	.338			8	.296
		4	.029			4	.043	6	14	2	.000			9	.484
		5	.095			5	.121			3	.001	7	12	2	.000
		6	.239			6	.296			4	.004			3	.000
		7	.455			7	.500			5	.017			4	.003
5	11	2	.000	6	8	2	.001			6	.058			5	.013
		3	.004			3	.005			7	.151			6	.046
		4	.022			4	.028			8	.299			7	.117
		5	.077			5	.086	7	7	2	.001			8	.247
		6	.201			6	.226			3	.004			9	.428
		7	.407			7	.413			4	.025	7	13	2	.000
5	12	2	.000	6	9	2	.000			5	.078			3	.000
		3	.003			3	.003			6	.209			4	.002
		4	.017			4	.019			7	.383			5	.010
		5	.063			5	.063	7	8	2	.000			6	.035
		6	.170			6	.175			3	.002			7	.095
		7	.365			7	.343			4	.015			8	.208
5	13	2	.000	6	10	2	.000			5	.051			9	.378
		3	.002			3	.002			6	.149	8	8	2	.000
		4	.013			4	.013			7	.296			3	.001
		5	.053			5	.047							4	.009
		6	.145			6	.137							5	.032
		7	.330			7	.287							6	.100
						8	.497							7	.214
														8	.405

(*Continued*)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table D (*Continued*)

Left-tail probabilities

n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P
8	9	2	.000	9	9	2	.000	10	10	2	.000	11	11	2	.000
		3	.001			3	.000			3	.000			3	.000
		4	.005			4	.003			4	.001			4	.000
		5	.020			5	.012			5	.004			5	.002
		6	.069			6	.044			6	.019			6	.007
		7	.157			7	.109			7	.051			7	.023
		8	.319			8	.238			8	.128			8	.063
		9	.500			9	.399			9	.242			9	.135
8	10	2	.000	9	10	2	.000			10	.414			10	.260
		3	.000			3	.000	10	11	2	.000			11	.410
		4	.003			4	.002			3	.000	11	12	2	.000
		5	.013			5	.008			4	.001			3	.000
		6	.048			6	.029			5	.003			4	.000
		7	.117			7	.077			6	.012			5	.001
		8	.251			8	.179			7	.035			6	.005
		9	.419			9	.319			8	.092			7	.015
8	11	2	.000	9	11	2	.000			9	.185			8	.044
		3	.000			3	.000			10	.335			9	.099
		4	.002			4	.001			11	.500			10	.202
		5	.009			5	.005	10	12	2	.000			11	.335
		6	.034			6	.020			3	.000	12	12	2	.000
		7	.088			7	.055			4	.000			3	.000
		8	.199			8	.135			5	.002			4	.000
		9	.352			9	.255			6	.008			5	.001
8	12	2	.000			10	.430			7	.024			6	.003
		3	.000	9	12	2	.000			8	.067			7	.009
		4	.001			3	.000			9	.142			8	.030
		5	.006			4	.001			10	.271			9	.070
		6	.025			5	.003			11	.425			10	.150
		7	.067			6	.014							11	.263
		8	.159			7	.040							12	.421
		9	.297			8	.103								
		10	.480			9	.205								
						10	.362								

(*Continued*)

Table D (*continued*)

							Right-tail probabilities								
n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P
2	2	4	.333	4	8	9	.071	5	11	11	.058	6	12	12	.075
2	3	5	.100			8	.212			10	.154			11	.217
		4	.500			7	.467			9	.374			10	.395
2	4	5	.200	4	9	9	.098	5	12	11	.075	6	13	13	.034
2	5	5	.286			8	.255			10	.181			12	.092
2	6	5	.357	4	10	9	.126	5	12	9	.421			11	.257
2	7	5	.417			8	.294	5	13	11	.092			10	.439
2	8	5	.467	4	11	9	.154			10	.208	6	14	13	.044
3	3	6	.100			8	.330			9	.465			12	.111
		5	.300	4	12	9	.181	5	14	11	.111			11	.295
3	4	7	.029			8	.363			10	.234			10	.480
		6	.200	4	13	9	.208	5	15	11	.129	7	7	14	.001
		5	.457			8	.393			10	.258			13	.004
3	5	7	.071	4	14	9	.234	6	6	12	.002			12	.025
		6	.286			8	.421			11	.013			11	.078
3	6	7	.119	4	15	9	.258			10	.067			10	.209
		6	.357			8	.446			9	.175			9	.383
3	7	7	.167	4	16	9	.282			8	.392	7	8	15	.000
		6	.417			8	.470	6	7	13	.001			14	.002
3	8	7	.212	5	5	10	.008			12	.008			13	.012
		6	.467			9	.040			11	.034			12	.051
3	9	7	.255			8	.167			10	.121			11	.133
3	10	7	.294			7	.357			9	.267			10	.296
3	11	7	.330	5	6	11	.002			8	.500			9	.486
3	12	7	.363			10	.024	6	8	13	.002	7	9	15	.001
3	13	7	.393			9	.089			12	.016			14	.006
3	14	7	.421			8	.262			11	.063			13	.025
3	15	7	.446			7	.478			10	.179			12	.084
3	16	7	.470	5	7	11	.008			9	.354			11	.194
3	17	7	.491			10	.045	6	9	13	.006			10	.378
4	4	8	.029			9	.146			12	.028	7	10	15	.002
		7	.114			8	.348			11	.098			14	.010
		6	.371	5	8	11	.016			10	.238			13	.043
4	5	9	.008			10	.071			9	.434			12	.121
		8	.071			9	.207	6	10	13	.010			11	.257
		7	.214			8	.424			12	.042			10	.451
		6	.500	5	9	11	.028			11	.136	7	11	15	.004
4	6	9	.024			10	.098			10	.294			14	.017
		8	.119			9	.266	6	11	13	.017			13	.064
		7	.310			8	.490			12	.058			12	.160
4	7	9	.045	5	10	11				11	.176			11	.318
		8	.167			10	.126			10	.346	7	12	15	.007
		7	.394			9	.322	6	12	13	.025			14	.025

(*Continued*)

Table D (*continued*)

						Right-tail probabilities					
n_1	n_2	R	P	n_1	n_2	R	P	n_1	n_2	R	P
7	12	13	.089	9	9	18	.00	10	13	13	.320
		12	.199			17	.000			12	.500
		11	.376			16	.003	10	12	21	.000
7	13	15	.010			15	.012			20	.000
		14	.034			14	.044			19	.001
		13	.116			13	.109			18	.006
		12	.238			12	.238			17	.020
		11	.430			11	.399			16	.056
8	8	16	.000	9	10	19	.000			15	.125
		15	.001			18	.000			14	.245
		14	.009			17	.001			13	.395
		13	.032			16	.008	11	11	22	.000
		12	.100			15	.026			21	.000
		11	.214			14	.077			20	.000
		10	.405			13	.166			19	.002
8	9	17	.000			12	.319			18	.007
		16	.001			11	.490			17	.023
		15	.004	9	11	19	.000			16	.063
		14	.020			18	.001			15	.135
		13	.061			17	.003			14	.260
		12	.157			16	.015			13	.410
		11	.298			15	.045	11	12	23	.000
		10	.500			14	.115			22	.000
8	10	17	.000			13	.227			21	.000
		16	.002			12	.395			20	.001
		15	.010	10	10	20	.000			19	.004
		14	.036			19	.000			18	.015
		13	.097			18	.000			17	.041
		12	.218			17	.001			16	.099
		11	.379			16	.004			15	.191
8	11	17	.001			17	.019			14	.335
		16	.004			16	.051			13	.493
		15	.018			15	.128	12	12	24	.000
		14	.057			14	.242			23	.000
		13	.138			13	.414			22	.000
		12	.278	10	11	21	.000			21	.001
		11	.453			20	.000			20	.003
8	12	17	.001			19	.000			19	.009
		16	.007			18	.003			18	.030
		15	.029			17	.010			17	.070
		14	.080			16	.035			16	.150
		13	.183			15	.085			15	.263
		12	.337			14	.185			14	.421

Source: Adapted from F. S. Swed and C. Eisenhart (1943), Tables for testing the randomness of grouping in a sequence of alternatives, *Annals of Mathematical Statistics*, 14, 66–87, with permission.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table E Runs Up and Down Distribution

Each table entry labeled P is the tail probability from each extreme to the value of R, the total number of runs up and down in a sequence of n observations, or equivalently, $n-1$ plus or minus signs.

n	*R*	*Left-tail P*	*R*	*Right-tail P*
3	1	.3333	2	.6667
4			3	.4167
	1	.0833	2	.9167
5	1	.0167	4	.2667
	2	.2500	3	.7500
6	1	.0028		
	2	.0861	5	.1694
	3	.4139	4	.5861
7	1	.0004	6	.1079
	2	.0250	5	.4417
	3	.1909	4	.8091
8	1	.0000		
	2	.0063	7	.0687
	3	.0749	6	.3250
	4	.3124	5	.6876
9	1	.0000		
	2	.0014		
	3	.0257	8	.0437
	4	.1500	7	.2347
	5	.4347	6	.5653
10	1	.0000		
	2	.0003	9	.0278
	3	.0079	8	.1671
	4	.0633	7	.4524
	5	.2427	6	.7573
11	1	.0000		
	2	.0001		
	3	.0022	10	.0177
	4	.0239	9	.1177
	5	.1196	8	.3540
	6	.3438	7	.6562
12	1	.0000		
	2	.0000		
	3	.0005		
	4	.0082	11	.0113
	5	.0529	10	.0821
	6	.1918	9	.2720
	7	.4453	8	.5547
13	1	.0000		
	2	.0000		
	3	.0001	12	.0072
	4	.0026	11	.0568
	5	.0213	10	.2058
	6	.0964	9	.4587
	7	.2749	8	.7251
14	1	.0000		
	2	.0000		
	3	.0000		
	4	.0007	13	.0046
	5	.0079	12	.0391
	6	.0441	11	.1536
	7	.1534	10	.3722
	8	.3633	9	.6367
15	1	.0000		
	2	.0000		
	3	.0000		
	4	.0002		
	5	.0027	14	.0029
	6	.0186	13	.0267
	7	.0782	12	.1134
	8	.2216	11	.2970
	9	.4520	10	.5480
16	1	.0000		
	2	.0000		
	3	.0000		
	4	.0001	15	.0019
	5	.0009	14	.0182
	6	.0072	13	.0828
	7	.0367	12	.2335
	8	.1238	11	.4631
	9	.2975	10	.7025

(*Continued*)

Table E (*Continued*)

n	*R*	*Left-tail P*	*R*	*Right-tail P*
17	1	.0000		
	2	.0000		
	3	.0000		
	4	.0000		
	5	.0003	16	.0012
	6	.0026	15	.0123
	7	.0160	14	.0600
	8	.0638	13	.1812
	9	.1799	12	.3850
	10	.3770	11	.6230
18	1	.0000		
	2	.0000		
	3	.0000		
	4	.0000		
	5	.0001		
	6	.0009	17	.0008
	7	.0065	16	.0083
	8	.0306	15	.0431
	9	.1006	14	.1389
	10	.2443	13	.3152
	11	.4568	12	.5432
19	1	.0000		
	2	.0000		
	3	.0000		
	4	.0000		
	4	.0000		
	5	.0000	18	.0005
	6	.0003	17	.0056
	7	.0025	16	.0308
	8	.0137	15	.1055
	9	.0523	14	.2546
	10	.1467	13	.4663
	11	.3144	12	.6856
20	1	.0000		
	2	.0000		
	3	.0000		
	4	.0000		
	5	.0000		
	6	.0001	19	.0003
	7	.0009	18	.0038
	8	.0058	17	.0218
	9	.0255	16	.0793
	10	.0821	15	.2031
	11	.2012	14	.3945
	12	.3873	13	.6127

n	*R*	*Left-tail P*	*R*	*Right-tail P*
21	1	.0000		
	2	.0000		
	3	.0000		
	4	.0000		
	5	.0000		
	6	.0000		
	7	.0003	20	.0002
	8	.0023	19	.0025
	9	.0117	18	.0154
	10	.0431	17	.0591
	11	.1202	16	.1602
	12	.2622	15	.3293
	13	.4603	14	.5397
22	1	.0000		
	2	.0000		
	3	.0000		
	4	.0000		
	5	.0000		
	6	.0000	21	.0001
	7	.0001	20	.0017
	8	.0009	19	.0108
	9	.0050	18	.0437
	10	.0213	17	.1251
	11	.0674	16	.2714
	12	.1661	15	.4688
	13	.3276	14	.6724
23	1	.0000		
	2	.0000		
	3	.0000		
	4	.0000		
	5	.0000		
	6	.0000		
	7	.0000	22	.0001
	8	.0003	21	.0011
	9	.0021	20	.0076
	10	.0099	19	.0321
	11	.0356	18	.0968
	12	.0988	17	.2211
	13	.2188	16	.4020
	14	.3953	15	.6047

(*Continued*)

Table E (*Continued*)

n	*R*	*Left-tail P*	*R*	*Right-tail P*	*n*	*R*	*Left-tail P*	*R*	*Right-tail P*
24	1	.0000			25	1	.0000		
	2	.0000				2	.0000		
	3	.0000				3	.0000		
	4	.0000				4	.0000		
	5	.0000				5	.0000		
	6	.0000				6	.0000		
	7	.0000				7	.0000	24	.0000
	8	.0001	23	.0000		8	.0000	23	.0005
	9	.0008	22	.0007		9	.0003	22	.0037
	10	.0044	21	.0053		10	.0018	21	.0170
	11	.0177	20	.0235		11	.0084	20	.0564
	12	.0554	19	.0742		12	.0294	19	.1423
	13	.1374	18	.1783		13	.0815	18	.2852
	14	.2768	17	.3405		14	.1827	17	.4708
	15	.4631	16	.5369		15	.3384	16	.6616

Source: Adapted from E. S. Edgington (1961), Probability table for number of runs of signs of first differences, *Journal of the American Statistical Association*, 56, 156–159, with permission.

Table F Kolmogorov-Smirnov One-Sample Statistic

Each table entry is the value of a Kolmogorov-Smirnov one-sample statistic D_n for sample size n such that its right-tail probability is the value given on the top row.

n	*.200*	*.100*	*.050*	*.020*	*.010*	n	*.200*	*.100*	*.050*	*.020*	*.010*
1	.900	.950	.975	.990	.995	21	.226	.259	.287	.321	.344
2	.684	.776	.842	.900	.929	22	.221	.253	.281	.314	.337
3	.565	.636	.780	.785	.829	23	.216	.247	.275	.307	.330
4	.493	.565	.624	.689	.734	24	.212	.242	.269	.301	.323
5	.447	.509	.563	.627	.669	25	.208	.238	.264	.295	.317
6	.410	.468	.519	.577	.617	26	.204	.233	.259	.290	.311
7	.381	.436	.483	.538	.576	27	.200	.229	.254	.284	.305
8	.358	.410	.454	.507	.542	28	.197	.225	.250	.279	.300
9	.339	.387	.430	.480	.513	29	.193	.221	.246	.275	.295
10	.323	.369	.409	.457	.489	30	.190	.218	.242	.270	.290
11	.308	.352	.391	.437	.468	31	.187	.214	.238	.266	.285
12	.296	.338	.375	.419	.449	32	.184	.211	.234	.262	.281
13	.285	.325	.361	.404	.432	33	.182	.208	.231	.258	.277
14	.275	.314	.349	.390	.418	34	.179	.205	.227	.254	.273
15	.266	.304	.338	.377	.404	35	.177	.202	.224	.251	.269
16	.258	.295	.327	.366	.392	36	.174	.199	.221	.247	.265
17	.250	.286	.318	.355	.381	37	.172	.196	.218	.244	.262
18	.244	.279	.309	.346	.371	38	.170	.194	.215	.241	.258
19	.237	.271	.301	.337	.361	39	.168	.191	.213	.238	.255
20	.232	.265	.294	.329	.352	40	.165	.189	.210	.235	.252

For $n > 40$, right-tail critical values based on the asymptotic distribution can be calculated as follows:

.200	.100	.050	.020	.010
$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

Source: Adapted from L. H. Miller (1956), Table of percentage points of Kolmogorov statistics, *Journal of the American Statistical Association*, **51**, 111–121, with permission.

Table G Binomial Distribution for $\theta = 0.5$

Each table entry labeled P is the tail probability from each extreme to the value of K, the number of successes in N Bernoulli trials with probability of success $\theta = 0.5$ on each trial.

N	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*
1	0	.5000	1	12	0	.0002	12	17	0	.0000	17
2	0	.2500	2		1	.0032	11		1	.0001	16
	1	.7500	1		2	.0193	10		2	.0012	15
3	0	.1250	3		3	.0730	9		3	.0064	14
	1	.5000	2		4	.1938	8		4	.0245	13
4	0	.0625	4		5	.3872	7		5	.0717	12
	1	.3125	3		6	.6128	6		6	.1662	11
	2	.6875	2	13	0	.0001	13		7	.3145	10
5	0	.0312	5		1	.0017	12		8	.5000	9
	1	.1875	4		2	.0112	11	18	0	.0000	18
	2	.5000	3		3	.0461	10		1	.0001	17
6	0	.0156	6		4	.1334	9		2	.0007	16
	1	.1094	5		5	.2905	8		3	.0038	15
	2	.3438	4		6	.5000	7		4	.0154	14
	3	.6562	3	14	0	.0000	14		5	.0481	13
7	0	.0078	7		1	.0009	13		6	.1189	12
	1	.0625	6		2	.0065	12		7	.2403	11
	2	.2266	5		3	.0287	11		8	.4073	10
	3	.5000	4		4	.0898	10		9	.5927	9
8	0	.0039	8		5	.2120	9	19	0	.0000	19
	1	.0352	7		6	.3953	8		1	.0000	18
	2	.1445	6		7	.6047	7		2	.0004	17
	3	.3633	5	15	0	.0000	15		3	.0022	16
	4	.6367	4		1	.0005	14		4	.0096	15
9	0	.0020	9		2	.0037	13		5	.0318	14
	1	.0195	8		3	.0176	12		6	.0835	13
	2	.0898	7		4	.0592	11		7	.1796	12
	3	.2539	6		5	.1509	10		8	.3238	11
	4	.5000	5		6	.3036	9		9	.5000	10
10	0	.0010	10		7	.5000	8	20	0	.0000	20
	1	.0107	9	16	0	.0000	16		1	.0000	19
	2	.0547	8		1	.0003	15		2	.0002	18
	3	.1719	7		2	.0021	14		3	.0013	17
	4	.3770	6		3	.0106	13		4	.0059	16
	5	.6230	5		4	.0384	12		5	.0207	15
11	0	.0005	11		5	.1051	11		6	.0577	14
	1	.0059	10		6	.2272	10		7	.1316	13
	2	.0327	9		7	.4018	9		8	.2517	12
	3	.1133	8		8	.5982	8		9	.4119	11
	4	.2744	7						10	.5881	10
	5	.5000	6								

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table H Probabilities for the Wilcoxon Signed-Rank Statistic
Each table entry labeled P is the tail probability from each extreme to the value of T, the Wilcoxon signed-rank statistic for sample side N, where T is interpreted as either T^+ or T^-.

N	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*
2	0	.250	3	7	0	.008	28	9	0	.002	45
	1	.500	2		1	.016	27		1	.004	44
3	0	.125	6		2	.023	26		2	.006	43
	1	.250	5		3	.039	25		3	.010	42
	2	.375	4		4	.055	24		4	.014	41
	3	.625	3		5	.078	23		5	.020	40
4	0	.062	10		6	.109	22		6	.027	39
	1	.125	9		7	.148	21		7	.037	38
	2	.188	8		8	.188	20		8	.049	37
	3	.312	7		9	.234	19		9	.064	36
	4	.438	6		10	.289	18		10	.082	35
	5	.562	5		11	.344	17		11	.102	34
5	0	.031	15		12	.406	16		12	.125	33
	1	.062	14		13	.469	15		13	.150	32
	2	.094	13		14	.531	14		14	.180	31
	3	.156	12	8	0	.004	36		15	.213	30
	4	.219	11		1	.008	35		16	.248	29
	5	.312	10		2	.012	34		17	.285	28
	6	.406	9		3	.020	33		18	.326	27
	7	.500	8		4	.027	32		19	.367	26
6	0	.016	21		5	.039	31		20	.410	25
	1	.031	20		6	.055	30		21	.455	24
	2	.047	19		7	.074	29		22	.500	23
	3	.078	18		8	.098	28	10	0	.001	55
	4	.109	17		9	.125	27		1	.002	54
	5	.156	16		10	.156	26		2	.003	53
	6	.219	15		11	.191	25		3	.005	52
	7	.281	14		12	.230	24		4	.007	51
	8	.344	13		13	.273	23		5	.010	50
	9	.422	12		14	.320	22		6	.014	49
	10	.500	11		15	.371	21		7	.019	48
					16	.422	20		8	.024	47
					17	.473	19		9	.032	46
					18	.527	18		10	.042	45

(*Continued*)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table H (*Continued*)

N	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*
10	11	.053	44	11	28	.350	38	13	0	.000	91
	12	.065	43		29	.382	37		1	.000	90
	13	.080	42		30	.416	36		2	.000	89
	14	.097	41		31	.449	35		3	.001	88
	15	.116	40		32	.483	34		4	.001	87
	16	.138	39		33	.517	33		5	.001	86
	17	.161	38	12	0	.000	78		6	.002	85
	18	.188	37		1	.000	77		7	.002	84
	19	.216	36		2	.001	76		8	.003	83
	20	.246	35		3	.001	75		9	.004	82
	21	.278	34		4	.002	74		10	.005	81
	22	.312	33		5	.002	73		11	.007	80
	23	.348	32		6	.003	72		12	.009	79
	24	.385	31		7	.005	71		13	.011	78
	25	.423	30		8	.006	70		14	.013	77
	26	.461	29		9	.008	69		15	.016	76
	27	.500	28		10	.010	68		16	.020	75
11	0	.000	66		11	.013	67		17	.024	74
	1	.001	65		12	.017	66		18	.029	73
	2	.001	64		13	.021	65		19	.034	72
	3	.002	63		14	.026	64		20	.040	71
	4	.003	62		15	.032	63		21	.047	70
	5	.005	61		16	.039	62		22	.055	69
	6	.007	60		17	.0456	61		23	.064	68
	7	.009	59		18	.055	60		24	.073	67
	8	.0125	58		19	.065	59		25	.084	66
	9	.016	57		20	.076	58		26	.095	65
	10	.021	56		21	.088	57		27	.108	64
	11	.027	55		22	.102	56		28	.122	63
	12	.034	54		23	.117	55		29	.137	62
	13	.042	53		24	.133	54		30	.153	61
	14	.051	52		25	.151	53		31	.170	60
	15	.062	51		26	.170	52		32	.188	59
	16	.074	50		27	.190	51		33	.207	58
	17	.087	49		28	.212	50		34	.227	57
	18	.103	48		29	.235	49		35	.249	56
	19	.120	47		30	.259	48		36	.271	55
	20	.139	46		31	.285	47		37	.294	54
	21	.160	45		32	.311	46		38	.318	53
	22	.183	44		33	.339	45		39	.342	52
	23	.207	43		34	.367	44		40	.368	51
	24	.232	42		35	.396	43		41	.393	50
	25	.260	41		36	.425	42		42	.420	49
	26	.289	40		37	.455	41		43	.446	48
	27	.319	39		38	.485	40		44	.473	47
					39	.515	39		45	.500	46

(*Continued*)

Table H (*Continued*)

N	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*	*N*	*Left tail*	*P*	*Right tail*
14	0	.000	105	14	46	.357	59	15	39	.126	81
	1	.000	104		47	.380	58		40	.138	80
	2	.000	103		48	.404	57		41	.151	79
	3	.000	102		49	.428	56		42	.165	78
	4	.000	101		50	.452	55		43	.180	77
	5	.001	100		51	.476	54		44	.195	76
	6	.001	99		52	.500	53		45	.211	75
	7	.001	98	15	0	.000	120		46	.227	74
	8	.002	97		1	.000	119		47	.244	73
	9	.002	96		2	.000	118		48	.262	72
	10	.003	95		3	.000	117		49	.281	71
	11	.003	94		4	.000	116		50	.300	70
	12	.004	93		5	.000	115		51	.319	69
	13	.005	92		6	.000	114		52	.339	68
	14	.007	91		7	.001	113		53	.360	67
	15	.008	90		8	.001	112		54	.381	66
	16	.010	89		9	.001	111		55	.402	65
	17	.012	88		10	.001	110		56	.423	64
	18	.0158	87		11	.002	109		57	.445	63
	19	.018	86		12	.002	108		58	.467	62
	20	.021	85		13	.003	107		59	.489	61
	21	.025	84		14	.003	106		60	.511	60
	22	.029	83		15	.004	105				
	23	.034	82		16	.005	104				
	24	.039	81		17	.006	103				
	25	.045	80		18	.008	102				
	26	.052	79		19	.009	101				
	27	.059	78		20	.011	100				
	28	.068	77		21	.013	99				
	29	.077	76		22	.015	98				
	30	.086	75		23	.018	97				
	31	.097	74		24	.021	96				
	32	.108	73		25	.024	95				
	33	.121	72		26	.028	94				
	34	.134	71		27	.032	93				
	35	.148	70		28	.036	92				
	36	.163	69		29	.042	91				
	37	.179	68		30	.047	90				
	38	.196	67		31	.053	89				
	39	.213	66		32	.060	88				
	40	.232	65		33	.068	87				
	41	.251	64		34	.076	86				
	42	.271	63		35	.084	85				
	43	.292	62		36	.094	84				
	44	.313	61		37	.104	83				
	45	.335	60		38	.115	82				

Source: Adapted from F. Wilcoxon, S.K. Katti, and R. A. Wilcox (1973), Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test, pp. 171–259 in Institute of Mathematical Statistics, ed., *Selected Tables in Mathematical Statistics* vol. I, American Mathematical Society, Providence, Rhode Island, with permission.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table I Kolmogorov-Smirnov Two-Sample Statistic

Each table entry labeled P is the right-tail probability of $mnD_{m,n}$, the Kolmogorov-Smirnov two-sample statistic for sample sizes m and n where $m \leqslant n$. The second portion of the table gives the value of $mnD_{m,n}$ such that its right-tail probability is the value given on the top row.

m	n	mnD	P	m	n	mnD	P	m	n	mnD	P
2	2	4	.333	3	6	18	.024	4	5	20	.016
2	3	6	.200			15	.095			16	.079
2	4	8	.133			12	.333			15	.143
2	5	10	.095	3	7	21	.017	4	6	24	.010
		8	.286			18	.067			20	.048
2	6	12	.071			15	.167			18	.095
		10	.214	3	8	24	.012			16	.181
2	7	14	.056			21	.048	4	7	28	.006
		12	.167			18	.121			24	.030
2	8	16	.044	3	9	27	.009			21	.067
		14	.133			24	.036			20	.121
2	9	18	.036			21	.091	4	8	32	.004
		16	.109			18	.236			28	.020
2	10	20	.030	3	10	30	.007			24	.085
		18	.091			27	.028			20	.222
		16	.182			24	.070	4	9	36	.003
2	11	22	.026			21	.140			32	.014
		20	.077	3	11	33	.005			28	.042
		18	.154			30	.022			27	.062
2	12	24	.022			27	.055			24	.115
		22	.066			24	.110	4	10	40	.002
		20	.132	3	12	36	.004			36	.010
3	3	9	.100			33	.018			32	.030
3	4	12	.057			30	.044			30	.046
		9	.229			27	.088			28	.084
3	5	15	.036			24	.189			26	.126
		12	.143	4	4	16	.029				
						12	.229				

(*Continued*)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table I (*Continued*)

m	n	mnD	P	m	n	mnD	P	m	n	mnD	P
4	11	44	.001	5	10	50	.001	6	10	60	.000
		40	.007			45	.004			54	.002
		36	.022			40	.019			50	.004
		33	.035			35	.061			48	.009
		32	.063			30	.166			44	.019
		29	.098	5	11	55	.000			42	.031
		28	.144			50	.003			40	.042
4	12	48	.001			45	.010			38	.066
		44	.005			44	.014			36	.092
		40	.016			40	.029			34	.125
		36	.048			39	.044	7	7	49	.001
		32	.112			35	.074			42	.008
5	5	25	.008			34	.106			35	.053
		20	.079	6	6	36	.002			28	.212
		15	.357			30	.026	7	8	56	.000
5	6	30	.004			24	.143			49	.002
		25	.026	6	7	42	.001			48	.005
		24	.048			36	.008			42	.013
		20	.108			35	.015			41	.024
5	7	35	.003			30	.038			40	.033
		30	.015			29	.068			35	.056
		28	.030			28	.091			34	.087
		25	.066			24	.147			33	.118
		23	.166	6	8	48	.001	7	9	63	.000
5	8	40	.022			42	.005			56	.001
		35	.009			40	.009			54	.003
		32	.020			36	.023			49	.008
		30	.042			34	.043			47	.015
		27	.079			32	.061			45	.021
		25	.126			30	.093			42	.034
5	9	45	.001			28	.139			40	.055
		40	.006	6	9	54	.000			38	.079
		36	.014			48	.003			36	.098
		35	.028			45	.006			35	.127
		31	.056			42	.014	8	8	64	.000
		30	.086			39	.028			56	.002
		27	.119			36	.061			48	.019
						33	.095			40	.087
						30	.176			32	.283

(*Continued*)

Table I (*Continued*)

$m=n$	*.200*	*.100*	*.050*	*.020*	*.010*
9	45	54	54	63	63
10	50	60	70	70	80
11	66	66	77	88	88
12	72	72	84	96	96
13	78	91	91	104	117
14	84	98	112	112	126
15	90	105	120	135	135
16	112	112	128	144	160
17	119	136	136	153	170
18	126	144	162	180	180
19	133	152	171	190	190
20	140	160	180	200	220

For m and n large, right-tail critical values based on the asymptotic distribution can be calculated as follows:

.200	.100	.050	.020	.010
$1.07\sqrt{N/mn}$	$1.22\sqrt{N/mn}$	$1.36\sqrt{N/mn}$	$1.52\sqrt{N/mn}$	$1.63\sqrt{N/mn}$

Source: Adapted from P. J. Kim and R. I. Jennrich (1973), Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov criterion $D_{mn}(m \leqslant n)$, pp. 79–170, in Institute of Mathematical Statistics, ed., *Selected Tables in Mathematical Statistics*, Vol. I, American Mathematical Society, Providence, Rhode Island, with permission.

Table J Probabilities for the Wilcoxon Rank-Sum Statistic
Each table entry labeled P is the tail probability from each extreme to the value of W_N, the Wilcoxon statistic for sample sizes m and n where $m \leqslant n$.

n	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*
	$m = 1$				$m = 2$				$m = 2$		
1	1	.500	2	2	3	.167	7	8	3	.022	19
2	1	.333	3		4	.333	6		4	.044	18
	2	.667	2		5	.667	5		5	.089	17
3	1	.250	4	3	3	.100	9		6	.133	16
	2	.500	3		4	.200	8		7	.200	15
4	1	.200	5		5	.400	7		8	.267	14
	2	.400	4		6	.600	6		9	.356	13
	3	.600	3	4	3	.067	11		10	.444	12
5	1	.167	6		4	.133	10		11	.556	11
	2	.333	5		5	.267	9	9	3	.018	21
	3	.500	4		6	.400	8		4	.036	20
6	1	.143	7		7	.600	7		5	.073	19
	2	.286	6	5	3	.048	13		6	.109	18
	3	.429	5		4	.095	12		7	.164	17
	4	.571	4		5	.190	11		8	.218	16
7	1	.125	8		6	.286	10		9	.291	15
	2	.250	7		7	.429	9		10	.364	14
	3	.375	6		8	.571	8		11	.455	13
	4	.500	5	6	3	.036	15		12	.545	12
8	1	.111	9		4	.071	14	10	3	.015	23
	2	.222	8		5	.143	13		4	.030	22
	3	.333	7		6	.214	12		5	.061	21
	4	.444	6		7	.321	11		6	.091	20
	5	.556	5		8	.429	10		7	.136	19
9	1	.100	10		9	.571	9		8	.182	18
	2	.200	9	7	3	.028	17		9	.242	17
	3	.300	8		4	.056	16		10	.303	16
	4	.400	7		5	.111	15		11	.379	15
	5	.500	6		6	.167	14		12	.455	14
10	1	.091	11		7	.250	13		13	.545	13
	2	.182	10		8	.333	12				
	3	.273	9		9	.444	11				
	4	.364	8		10	.556	10				
	5	.455	7								
	6	.545	6								

(*Continued*)

Table J (*Continued*)

n	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*
		$m = 3$				$m = 3$				$m = 4$	
3	6	.050	15	8	6	.006	30	4	10	.014	26
	7	.100	14		7	.012	29		11	.029	25
	8	.200	13		8	.024	28		12	.057	24
	9	.350	12		9	.042	27		13	.100	23
	10	.500	11		10	.067	26		14	.171	22
4	6	.029	18		11	.097	25		15	.243	21
	7	.057	17		12	.139	24		16	.343	20
	8	.114	16		13	.188	23		17	.443	19
	9	.200	15		14	.248	22		18	.557	18
	10	.314	14		15	.315	21	5	10	.008	30
	11	.429	13		16	.388	20		11	.016	29
	12	.571	12		17	.461	19		12	.032	28
5	6	.018	21		18	.539	18		13	.056	27
	7	.036	20	9	6	.005	33		14	.095	26
	8	.071	19		7	.009	32		15	.143	25
	9	.125	18		8	.018	31		16	.206	24
	10	.196	17		9	.032	30		17	.278	23
	11	.286	16		10	.050	29		18	.365	22
	12	.393	15		11	.073	28		19	.452	21
	13	.500	14		12	.105	27		20	.548	20
6	6	.012	24		13	.141	26	6	10	.005	34
	7	.024	23		14	.186	25		11	.010	33
	8	.048	22		15	.241	24		12	.019	32
	9	.083	21		16	.300	23		13	.033	31
	10	.131	20		17	.364	22		14	.057	30
	11	.190	19		18	.432	21		15	.086	29
	12	.274	18		19	.500	20		16	.129	28
	13	.357	17	10	6	.003	36		17	.176	27
	14	.452	16		7	.007	35		18	.238	26
	15	.548	15		8	.014	34		19	.305	25
7	6	.008	27		9	.024	33		20	.381	24
	7	.017	26		10	.038	32		21	.457	23
	8	.033	25		11	.056	31		22	.543	22
	9	.058	24		12	.080	30				
	10	.092	23		13	.108	29				
	11	.133	22		14	.143	28				
	12	.192	21		15	.185	27				
	13	.258	20		16	.234	26				
	14	.333	19		17	.287	25				
	15	.417	18		18	.346	24				
	16	.500	17		19	.406	23				
					20	.469	22				
					21	.531	21				

(*Continued*)

Table J (*Continued*)

n	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*
		$m=4$				$m=4$				$m=5$	
7	10	.003	38	9	10	.001	46	5	15	.004	40
	11	.006	37		11	.003	45		16	.008	39
	12	.012	36		12	.006	44		17	.016	38
	13	.021	35		13	.010	43		18	.028	37
	14	.036	34		14	.017	42		19	.048	36
	15	.055	33		15	.025	41		20	.075	35
	16	.082	32		16	.038	40		21	.111	34
	17	.115	31		17	.053	39		22	.155	33
	18	.158	30		18	.074	38		23	.210	32
	19	.206	29		19	.099	37		24	.274	31
	20	.264	28		20	.130	36		25	.345	30
	21	.324	27		21	.165	35		26	.421	29
	22	.394	26		22	.207	34		27	.500	28
	23	.464	25		23	.252	33	6	15	.002	45
	24	.536	24		24	.302	32		16	.004	44
8	10	.002	42		25	.355	31		17	.009	43
	11	.004	41		26	.413	30		18	.015	42
	12	.008	40		27	.470	29		19	.026	41
	13	.014	39		28	.530	28		20	.041	40
	14	.024	38	10	10	.001	50		21	.063	39
	15	.036	37		11	.002	49		22	.089	38
	16	.055	36		12	.004	48		23	.123	37
	17	.077	35		13	.007	47		24	.165	36
	18	.107	34		14	.012	46		25	.214	35
	19	.141	33		15	.018	45		26	.268	34
	20	.184	32		16	.027	44		27	.331	33
	21	.230	31		17	.038	43		28	.396	32
	22	.285	30		18	.053	42		29	.465	31
	23	.341	29		19	.071	41		30	.535	30
	24	.404	28		20	.094	40				
	25	.467	27		21	.120	39				
	26	.533	26		22	.152	38				
					23	.187	37				
					24	.227	36				
					25	.270	35				
					26	.318	34				
					27	.367	33				
					28	.420	32				
					29	.473	31				
					30	.527	30				

(*Continued*)

Table J (*Continued*)

n	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*
	$m=5$				$m=5$				$m=6$		
7	15	.001	50	9	15	.000	60	6	21	.001	57
	16	.003	49		16	.001	59		22	.002	56
	17	.005	48		17	.002	58		23	.004	55
	18	.009	47		18	.003	57		24	.008	54
	19	.015	46		19	.006	56		25	.013	53
	20	.024	45		20	.009	55		26	.021	52
	21	.037	44		21	.014	54		27	.032	51
	22	.053	43		22	.021	53		28	.047	50
	23	.074	42		23	.030	52		29	.066	49
	24	.101	41		24	.041	51		30	.090	48
	25	.134	40		25	.056	50		31	.120	47
	26	.172	39		26	.073	49		32	.155	46
	27	.216	38		27	.095	48		33	.197	45
	28	.265	37		28	.120	47		34	.242	44
	29	.319	36		29	.149	46		35	.294	43
	30	.378	35		30	.182	45		36	.350	42
	31	.438	34		31	.219	44		37	.409	41
	32	.500	33		32	.259	43		38	.469	40
8	15	.001	55		33	.303	42		39	.531	39
	16	.002	54		34	.350	41	7	21	.001	63
	17	.003	53		35	.399	40		22	.001	62
	18	.005	52		36	.449	39		23	.002	61
	19	.009	51		37	.500	38		24	.004	60
	20	.015	50	10	15	.000	65		25	.007	59
	21	.023	49		16	.001	64		26	.011	58
	22	.033	48		17	.001	63		27	.017	57
	23	.047	47		18	.002	62		28	.026	56
	24	.064	46		19	.004	61		29	.037	55
	25	.085	45		20	.006	60		30	.051	54
	26	.111	44		21	.010	59		31	.069	53
	27	.142	43		22	.014	58		32	.090	52
	28	.177	42		23	.020	57		33	.117	51
	29	.218	41		24	.028	56		34	.147	50
	30	.262	40		25	.038	55		35	.183	49
	31	.311	39		26	.050	54		36	.223	48
	32	.362	38		27	.065	53		37	.267	47
	33	.416	37		28	.082	52		38	.314	46
	34	.472	36		29	.103	51		39	.365	45
	35	.528	35		30	.127	50		40	.418	44
					31	.155	49		41	.473	43
					32	.0185	48		42	.527	42
					33	.220	47				
					34	.257	46				
					35	.297	45				
					36	.339	44				
					37	.384	43				
					38	.430	42				
					39	.477	41				
					40	.523	40				

(*Continued*)

Table J (*Continued*)

n	Left tail	P	Right tail	n	Left tail	P	Right tail	n	Left tail	P	Right tail
	$m=6$				$m=6$				$m=7$		
8	21	.000	69	9	41	.228	55	7	28	.000	77
	22	.001	68		42	.264	54		29	.001	76
	23	.001	67		43	.303	53		30	.001	75
	24	.002	66		44	.344	52		31	.002	74
	25	.004	65		45	.388	51		32	.003	73
	26	.006	64		46	.432	50		33	.006	72
	27	.010	63		47	.477	49		34	.009	71
	28	.015	62		48	.523	48		35	.013	70
	29	.021	61	10	21	.000	81		36	.019	69
	30	.030	60		22	.000	80		37	.027	68
	31	.041	59		23	.000	79		38	.036	67
	32	.054	58		24	.001	78		39	.049	66
	33	.071	57		25	.001	77		40	.064	65
	34	.091	56		26	.002	76		41	.082	64
	35	.114	55		27	.004	75		42	.104	63
	36	.141	54		28	.005	74		43	.130	62
	37	.172	53		29	.008	73		44	.159	61
	38	.207	52		30	.011	72		45	.191	60
	39	.245	51		31	.016	71		46	.228	59
	40	.286	50		32	.021	70		47	.267	58
	41	.331	49		33	.028	69		48	.310	57
	42	.377	48		34	.036	68		49	.355	56
	43	.426	47		35	.047	67		50	.402	55
	44	.475	46		36	.059	66		51	.451	54
	45	.525	45		37	.074	65		52	.500	53
9	21	.000	75		38	.090	64	8	28	.000	84
	22	.000	74		39	.110	63		29	.000	83
	23	.001	73		40	.132	62		30	.001	82
	24	.001	72		41	.157	61		31	.001	81
	25	.002	71		42	.184	60		32	.002	80
	26	.004	70		43	.214	59		33	.003	79
	27	.006	69		44	.246	58		34	.005	78
	28	.009	68		45	.281	57		35	.007	77
	29	.013	67		46	.318	56		36	.010	76
	30	.018	66		47	.356	55		37	.014	75
	31	.025	65		48	.396	54		38	.020	74
	32	.033	64		49	.437	53		39	.027	73
	33	.044	63		50	.479	52		40	.036	72
	34	.057	62		51	.521	51		41	.047	71
	35	.072	61						42	.060	70
	36	.091	60						43	.076	69
	37	.112	59						44	.095	68
	38	.136	58						45	.116	67
	39	.164	57						46	.140	66
	40	.194	56						47	.168	65

(*Continued*)

Table J (*Continued*)

n	Left tail	P	Right tail	n	Left tail	P	Right tail	n	Left tail	P	Right tail
	$m = 7$				$m = 7$				$m = 8$		
8	48	.198	64	10	28	.000	98	8	36	.000	100
	49	.232	63		29	.000	97		37	.000	99
	50	.268	62		30	.000	96		38	.000	98
	51	.306	61		31	.000	95		39	.001	97
	52	.347	60		32	.001	94		40	.001	96
	53	.389	59		33	.001	93		41	.001	95
	54	.433	58		34	.002	92		42	.002	94
	55	.478	57		35	.002	91		43	.003	93
	56	.522	56		36	.003	90		44	.005	92
9	28	.000	91		37	.005	89		45	.007	91
	29	.000	90		38	.007	88		46	.010	90
	30	.000	89		39	.009	87		47	.014	89
	31	.001	88		40	.012	86		48	.019	88
	32	.001	87		41	.017	85		49	.025	87
	33	.002	86		42	.022	84		50	.032	86
	34	.003	85		43	.028	83		51	.041	85
	35	.004	84		44	.035	82		52	.052	84
	36	.006	83		45	.044	81		53	.065	83
	37	.008	82		46	.054	80		54	.080	82
	38	.011	81		47	.067	79		55	.097	81
	39	.016	80		48	.081	78		56	.117	80
	40	.021	79		49	.097	77		57	.139	79
	41	.027	78		50	.115	76		58	.164	78
	42	.036	77		51	.135	75		59	.191	77
	43	.045	76		52	.157	74		60	.221	76
	44	.057	75		53	.182	73		61	.253	75
	45	.071	74		54	.209	72		62	.287	74
	46	.087	73		55	.237	71		63	.323	73
	47	.105	72		56	.268	70		64	.360	72
	48	.126	71		57	.300	69		65	.399	71
	49	.150	70		58	.335	68		66	.439	70
	50	.176	69		59	.370	67		67	.480	69
	51	.204	68		60	.406	66		68	.520	68
	52	.235	67		61	.443	65	9	36	.000	108
	53	.268	66		62	.481	64		37	.000	107
	54	.303	65		63	.519	63		38	.000	106
	55	.340	64						39	.000	105
	56	.379	63						40	.000	104
	57	.419	62						41	.001	103
	58	.459	61						42	.001	102
	59	.500	60						43	.002	101

(*Continued*)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table J (*Continued*)

n	Left tail	P	Right tail	n	Left tail	P	Right tail	n	Left tail	P	Right tail
		$m = 8$				$m = 8$				$m = 9$	
9	44	.003	100	10	36	.000	116	9	45	.000	126
	45	.004	99		37	.000	115		46	.000	125
	46	.006	98		38	.000	114		47	.000	124
	47	.008	97		39	.000	113		48	.000	123
	48	.010	96		40	.000	112		49	.000	122
	49	.014	95		41	.000	111		50	.000	121
	50	.018	94		42	.001	110		51	.001	120
	51	.023	93		43	.001	109		52	.001	119
	52	.030	92		44	.002	108		53	.001	118
	53	.037	91		45	.002	107		54	.002	117
	54	.046	90		46	.003	106		55	.003	116
	55	.057	89		47	.004	105		56	.004	115
	56	.069	88		48	.006	104		57	.005	114
	57	.084	87		49	.008	103		58	.007	113
	58	.100	86		50	.010	102		59	.009	112
	59	.118	85		51	.013	101		60	.012	111
	60	.138	84		52	.017	100		61	.016	110
	61	.161	83		53	.022	99		62	.020	109
	62	.185	82		54	.027	98		63	.025	108
	63	.212	81		55	.034	97		64	.031	107
	64	.240	80		56	.042	96		65	.039	106
	65	.271	79		57	.051	95		66	.047	105
	66	.303	78		58	.061	94		67	.057	104
	67	.336	77		59	.073	93		68	.068	103
	68	.371	76		60	.086	92		69	.081	102
	69	.407	75		61	.102	91		70	.095	101
	70	.444	74		62	.118	90		71	.111	100
	71	.481	73		63	.137	89		72	.129	99
	72	.519	72		64	.158	88		73	.149	98
					65	.180	87		74	.170	97
					66	.204	86		75	.193	96
					67	.230	85		76	.218	95
					68	.257	84		77	.245	94
					69	.286	83		78	.273	93
					70	.317	82		79	.302	92
					71	.348	81		80	.333	91
					72	.381	80		81	.365	90
					73	.414	79		82	.398	89
					74	.448	78		83	.432	88
					75	.483	77		84	.466	87
					76	.517	76		85	.500	86

(*Continued*)

Table J (*Continued*)

n	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*	*n*	*Left tail*	*P*	*Right tail*
	$m = 9$				$m = 9$				$m = 10$		
10	45	.000	135	10	78	.178	102	10	73	.007	137
	46	.000	134		79	.200	101		74	.009	136
	47	.000	133		80	.223	100		75	.012	135
	48	.000	132		81	.248	99		76	.014	134
	49	.000	131		82	.274	98		77	.018	133
	50	.000	130		83	.302	97		78	.022	132
	51	.000	129		84	.330	96		79	.026	131
	52	.000	128		85	.360	95		80	.032	130
	53	.001	127		86	.390	94		81	.038	129
	54	.001	126		87	.421	93		82	.045	128
	55	.001	125		88	.452	92		83	.053	127
	56	.002	124		89	.484	91		84	.062	126
	57	.003	123		90	.516	90		85	.072	125
	58	.004	122						86	.083	124
	59	.005	121						87	.095	123
	60	.007	120		$m = 10$				88	.109	122
	61	.009	119						89	.124	121
	62	.011	118	10	55	.000	155		90	.140	120
	63	.014	117		56	.000	154		91	.157	119
	64	.017	116		57	.000	153		92	.176	118
	65	.022	115		58	.000	152		93	.197	117
	66	.027	114		59	.000	151		94	.218	116
	67	.033	113		60	.000	150		95	.241	115
	68	.039	112		61	.000	149		96	.264	114
	69	.047	111		62	.000	148		97	.289	113
	70	.056	110		63	.000	147		98	.315	112
	71	.067	109		64	.001	146		99	.342	111
	72	.078	108		65	.001	145		100	.370	110
	73	.091	107		66	.001	144		101	.398	109
	74	.106	106		67	.001	143		102	.427	108
	75	.121	105		68	.002	142		103	.456	107
	76	.139	104		69	.003	141		104	.485	106
	77	.158	103		70	.003	140		105	.515	105
					71	.004	139				
					72	.006	138				

Source: Adapted from F. Wilcoxon, S. K. Katti, and R. A. Wilcox (1973), Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test, pp. 172–259, in Institute of Mathematical Statistics, ed., *Selected Tables in Mathematical Statistics*, vol. I, American Mathematical Society, Providence, Rhode Island, with permission.

Table K Kruskal-Wallis Test Statistic

Each table entry is the smallest value of the Kruskal-Wallis H such that its right-tail probability is less than or equal to the value given on the top row for $k=3$, each sample size less than or equal to 5.

	Right-tail probability for H				
n_1, n_2, n_3	*0.100*	*0.050*	*0.020*	*0.010*	*0.001*
2, 2, 2	4.571	—	—	—	—
3, 2, 1	4.286	—	—	—	—
3, 2, 2	4.500	4.714	—	—	—
3, 3, 1	4.571	5.143	—	—	—
3, 3, 2	4.556	5.361	6.250	—	—
3, 3, 3	4.622	5.600	6.489	7.200	—
4, 2, 1	4.500	—	—	—	—
4, 2, 2	4.458	5.333	6.000	—	—
4, 3, 1	4.056	5.208	—	—	—
4, 3, 2	4.511	5.444	6.144	6.444	—
4, 3, 3	4.709	5.791	6.564	6.745	—
4, 4, 1	4.167	4.967	6.667	6.667	—
4, 4, 2	4.555	5.455	6.600	7.036	—
4, 4, 3	4.545	5.598	6.712	7.144	8.909
4, 4, 4	4.654	5.692	6.962	7.654	9.269
5, 2, 1	4.200	5.000	—	—	—
5, 2, 2	4.373	5.160	6.000	6.533	—
5, 3, 1	4.018	4.960	6.044	—	—
5, 3, 2	4.651	5.251	6.124	6.909	—
5, 3, 3	4.533	5.648	6.533	7.079	8.727
5, 4, 1	3.987	4.985	6.431	6.955	—
5, 4, 2	4.541	5.273	6.505	7.205	8.591
5, 4, 3	4.549	5.656	6.676	7.445	8.795
5, 4, 4	4.668	5.657	6.953	7.760	9.168
5, 5, 1	4.109	5.127	6.145	7.309	—
5, 5, 2	4.623	5.338	6.446	7.338	8.938
5, 5, 3	4.545	5.705	6.866	7.578	9.284
5, 5, 4	4.523	5.666	7.000	7.823	9.606
5, 5, 5	4.560	5.780	7.220	8.000	9.920

For $k>3$, right-tail probabilities on H are found from Table B with $k-1$ degrees of freedom.

Source: Adapted from R. L. Iman, D. Quade, and D. A. Alexander (1975), Exact probability levels for the Kruskal-Wallis test, pp. 329–384, in Institute of Mathematical Statistics ed., *Selected Tables in Mathematical Statistics*, vol. III, American Mathematical Society, Providence, Rhode Island, with permission.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table L Kendall's Tau Statistic

Each table entry labelled P is the right-tail probability for T, the Kendall tau statistic for sample size n, and also the left-tail probability for $-T$. The second portion of the table gives the value of $T(-T)$ such that its right (left-tail) probability is the value given in the top row.

n	T	P	n	T	P	n	T	P	n	T	P
3	1.000	.167	7	1.000	.000	9	1.000	.000	10	1.000	.000
	.333	.500		.905	.001		.944	.000		.956	.000
4	1.000	.042		.810	.005		.889	.000		.911	.000
	.667	.167		.714	.015		.833	.000		.867	.000
	.333	.375		.619	.035		.778	.001		.822	.000
	.000	.625		.524	.068		.722	.003		.778	.000
5	1.000	.008		.429	.119		.667	.006		.733	.001
	.800	.042		.333	.191		.611	.012		.689	.002
	.600	.117		.238	.281		.556	.022		.644	.005
	.400	.242		.143	.386		.500	.038		.600	.008
	.200	.408		.048	.500		.444	.060		.556	.014
	.000	.592	8	1.000	.000		.389	.090		.511	.023
6	1.000	.001		.929	.000		.333	.130		.467	.036
	.867	.008		.857	.001		.278	.179		.422	.054
	.733	.028		.786	.003		.222	.238		.378	.078
	.600	.068		.714	.007		.167	.306		.333	.108
	.467	.136		.643	.016		.111	.381		.289	.146
	.333	.235		.571	.031		.056	.460		.244	.190
	.200	.360		.500	.054		.000	.540		.200	.242
	.067	.500		.429	.089					.156	.300
				.357	.138					.111	.364
				.286	.199					.067	.431
				.214	.274					.022	.500
				.143	.360						
				.071	.452						
				.000	.548						

(*Continued*)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table L (*Continued*)

n	*.100*	*.050*	*.025*	*.010*	*.005*
11	.345	.418	.491	.564	.600
12	.303	.394	.455	.545	.576
13	.308	.359	.436	.513	.564
14	.275	.363	.407	.473	.516
15	.276	.333	.390	.467	.505
16	.250	.317	.383	.433	.483
17	.250	.309	.368	.426	.471
18	.242	.294	.346	.412	.451
19	.228	.287	.333	.392	.439
20	.221	.274	.326	.379	.421
21	.210	.267	.314	.371	.410
22	.203	.264	.307	.359	.394
23	.202	.257	.296	.352	.391
24	.196	.246	.290	.341	.377
25	.193	.240	.287	.333	.367
26	.188	.237	.280	.329	.360
27	.179	.231	.271	.322	.356
28	.180	.228	.265	.312	.344
29	.172	.222	.261	.310	.340
30	.172	.218	.255	.301	.333

Source: The tail probabilities ($n \leqslant 10$) are adapted from M. G. Kendall (1948, 4th ed. 1970), *Rank Correlation Methods*, Charles Griffin & Co., Ltd., London and High Wycombe, with permission. The quantiles ($11 \leqslant n \leqslant 30$) are adapted from L. Kaarsemaker and A. van Wijngaarden (1953), Tables for use in rank correlation, *Statistica Neerlandica*, **7**, 41–54, with permission.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table M Spearman's Coefficient of Rank Correlation
Each table entry labeled P is the right-tail probability for R, Spearman's coefficient of rank correlation for sample size n, and also the left-tail probability for $-R$. The second portion of he table gives the value of $R(-R)$ such that its right-tail (left-tail) probability is the value given on the top row.

n	R	P	n	R	P	n	R	P	n	R	P
3	1.000	.167	7	1.000	.000	8	.810	.011	9	1.000	.000
	.500	.500		.964	.001		.786	.014		.983	.000
4	1.000	.042		.929	.003		.762	.018		.967	.000
	.800	.167		.893	.006		.738	.023		.950	.000
	.600	.208		.857	.012		.714	.029		.933	.000
	.400	.375		.821	.017		.690	.035		.917	.001
	.200	.458		.786	.024		.667	.042		.900	.001
	.000	.542		.750	.033		.643	.048		.883	.002
5	1.000	.008		.714	.044		.619	.057		.867	.002
	.900	.042		.679	.055		.595	.066		.850	.003
	.800	.067		.643	.069		.571	.076		.833	.004
	.700	.117		.607	.083		.548	.085		.817	.005
	.600	.175		.571	.100		.524	.098		.800	.007
	.500	.225		.536	.118		.500	.108		.783	.009
	.400	.258		.500	.133		.476	.122		.767	.011
	.300	.342		.464	.151		.452	.134		.750	.013
	.200	.392		.429	.177		.429	.150		.733	.016
	.100	.475		.393	.198		.405	.163		.717	.018
	.000	.525		.357	.222		.381	.180		.700	.022
6	1.000	.001		.321	.249		.357	.195		.683	.025
	.943	.008		.286	.278		.333	.214		.667	.029
	.886	.017		.250	.297		.310	.231		.650	.033
	.829	.029		.214	.331		.286	.250		.633	.038
	.771	.051		.179	.357		.262	.268		.617	.043
	.714	.068		.143	.391		.238	.291		.600	.048
	.657	.088		.107	.420		.214	.310		.583	.054
	.600	.121		.071	.453		.190	.332		.567	.060
	.543	.149		.036	.482		.167	.352		.550	.066
	.486	.178		.000	.518		.143	.376		.533	.074
	.429	.210	8	1.000	.000		.119	.397		.517	.081
	.371	.249		.976	.000		.095	.420		.500	.089
	.314	.282		.952	.001		.071	.441		.483	.097
	.257	.329		.929	.001		.048	.467		.467	.106
	.200	.357		.905	.002		.024	.488		.450	.115
	.143	.401		.881	.004		.000	.512		.433	.125
	.086	.460		.857	.005					.417	.135
	.029	.500		.833	.008					.400	.146

(*Continued*)

Table M (*Continued*)

n	*R*	*P*	*n*	*R*	*P*	*n*	*R*	*P*	*n*	*R*	*P*
9	.383	.156	10	.964	.000	10	.636	.027	10	.309	.193
	.367	.168		.952	.000		.624	.030		.297	.203
	.350	.179		.939	.000		.612	.033		.285	.214
	.333	.193		.927	.000		.600	.037		.273	.224
	.317	.205		.915	.000		.588	.040		.261	.235
	.300	.218		.903	.000		.576	.044		.248	.246
	.283	.231		.891	.001		.564	.048		.236	.257
	.267	.247		.879	.001		.552	.052		.224	.268
	.250	.260		.867	.001		.539	.057		.212	.280
	.233	.276		.855	.001		.527	.062		.200	.292
	.217	.290		.842	.002		.515	.067		.188	.304
	.200	.307		.830	.002		.503	.072		.176	.316
	.183	.322		.818	.003		.491	.077		.164	.328
	.167	.339		.806	.004		.479	.083		.152	.341
	.150	.354		.794	.004		.467	.089		.139	.354
	.133	.372		.782	.005		.455	.096		.127	.367
	.117	.388		.770	.007		.442	.102		.115	.379
	.100	.405		.758	.008		.430	.109		.103	.393
	.083	.422		.745	.009		.418	.116		.091	.406
	.067	.440		.733	.010		.406	.124		.079	.419
	.050	.456		.721	.012		.394	.132		.067	.433
	.033	.474		.709	.013		.382	.139		.055	.446
	.017	.491		.697	.015		.370	.148		.042	.459
	.000	.509		.685	.017		.358	.156		.030	.473
10	1.000	.000		.673	.019		.345	.165		.018	.486
	.988	.000		.661	.022		.333	.174		.006	.500
	.976	.000		.648	.025		.321	.184			

(*Continued*)

Table M (*Continued*)

n	*.100*	*.050*	*.025*	*.010*	*.005*	*.001*
11	.427	.536	.618	.709	.764	.855
12	.406	.503	.587	.678	.734	.825
13	.385	.484	.560	.648	.703	.797
14	.367	.464	.538	.626	.679	.771
15	.354	.446	.521	.604	.657	.750
16	.341	.429	.503	.585	.635	.729
17	.329	.414	.488	.566	.618	.711
18	.317	.401	.474	.550	.600	.692
19	.309	.391	.460	.535	.584	.675
20	.299	.380	.447	.522	.570	.660
21	.292	.370	.436	.509	.556	.647
22	.284	.361	.425	.497	.544	.633
23	.278	.353	.416	.486	.532	.620
24	.275	.344	.407	.476	.521	.608
25	.265	.337	.398	.466	.511	.597
26	.260	.331	.390	.457	.501	.586
27	.255	.324	.383	.449	.492	.576
28	.250	.318	.376	.441	.483	.567
29	.245	.312	.369	.433	.475	.557
30	.241	.307	.363	.426	.467	.548

Source: The tail probabilities ($n \leqslant 10$) are adapted from M. G. Kendall (1948, 4th ed. 1970), *Rank Correlation Methods,* Charles Griffin & Co., Ltd., London and High Wycombe, with permission. The quantiles ($11 \leqslant n \leqslant 30$) are adapted from G. J. Glasser and R. F. Winter (1961), Critical values of the rank correlation coefficient for testing the hypothesis of independence, *Biometrika*, **48**, 444–448, with permission of the Biometrika Trustees and the authors.

Table N Friedman's Analysis-of-Variance Statistic and Kendall's Coefficient of Concordance

Each table entry labeled P is the right-tail probability for the sum of squares S used in Friedman's analysis-of-variance statistic with n treatments and k blocks and in Kendall's coefficient of concordance with k sets of ranking of n objects.

n	k	S	P	n	k	S	P	n	k	S	P	n	k	S	P
3	2	8	.167	3	7	98	.000	4	2	20	.042	4	4	80	.000
		6	.500			96	.000			18	.167			78	.001
	3	18	.028			86	.000			16	.208			76	.001
		14	.194			78	.001			14	.375			74	.001
		8	.361			74	.003			12	.458			72	.002
	4	32	.005			72	.004		3	45	.002			70	.003
		26	.042			62	.008			43	.002			68	.003
		24	.069			56	.016			41	.017			66	.006
		18	.125			54	.021			37	.033			64	.007
		14	.273			50	.027			35	.054			62	.012
		8	.431			42	.051			33	.075			58	.014
	5	50	.001			38	.085			29	.148			56	.019
		42	.008			32	.112			27	.175			54	.033
		38	.024			26	.192			25	.207			52	.036
		32	.039			24	.237			21	.300			50	.052
		26	.093			18	.305			19	.342			48	.054
		24	.124			14	.486			17	.446			46	.068
		18	.182		8	128	.000							44	.077
		14	.367			126	.000							42	.094
	6	72	.000			122	.000							40	.105
		62	.002			114	.000							38	.141
		56	.006			104	.000							36	.158
		54	.008			98	.001							34	.190
		50	.012			96	.001							32	.200
		42	.029			86	.002							30	.242
		38	.052			78	.005							26	.324
		32	.072			74	.008							24	.355
		26	.142			72	.010							22	.389
		24	.184			62	.018							20	.432
		18	.252			56	.030								
		14	.430			54	.038								
						50	.047								
						42	.079								
						38	.120								
						32	.149								
						26	.236								
						24	.285								
						18	.355								

Source: Adapted from M. G. Kendall (1948, 4th ed. 1970), *Rank Correlation Methods*, Charles Griffin & Co., Ltd., London and High Wycombe, with permission.

Table O Lilliefors's Test for Normal Distribution Critical Values
Table entries for any sample size N are the values of a Lilliefors's random variable with right-tail probability as given in the top row.

Sample Size (N)	*Significance level*			
	0.100	*0.05*	*0.010*	*0.001*
4	.344	.375	.414	.432
5	.320	.344	.398	.427
6	.298	.323	.369	.421
7	.281	.305	.351	.399
8	.266	.289	.334	.383
9	.252	.273	.316	.366
10	.240	.261	.305	.350
11	.231	.251	.291	.331
12	.223	.242	.281	.327
14	.208	.226	.262	.302
16	.195	.213	.249	.291
18	.185	.201	.234	.272
20	.176	.192	.223	.266
25	.159	.173	.202	.236
30	.146	.159	.186	.219
40	.127	.139	.161	.190
50	.114	.125	.145	.173
60	.105	.114	.133	.159
75	.094	.102	.119	.138
100	.082	.089	.104	.121
Over 100	$.816/\sqrt{N}$	$.888/\sqrt{N}$	$1.038/\sqrt{N}$	$1.212/\sqrt{N}$

Source: Adapted from R. L. Edgeman and R. C. Scott (1987), Lilliefors's tests for transformed variables, *Brazilian Journal of Probability and Statistics*, **1**, 101–112, with permission.

Table P Significance Points of $T_{XY.Z}$ (*for Kendall's Partial Rank-Correlation Coefficient*)

	One-tailed level of significance			
m	*0.005*	*0.01*	*0.025*	*0.05*
3	1	1	1	1
4	1	1	1	0.707
5	1	0.816	0.802	0.667
6	0.866	0.764	0.667	0.600
7	0.761	0.712	0.617	0.527
8	0.713	0.648	0.565	0.484
9	0.660	0.602	0.515	0.443
10	0.614	0.562	0.480	0.413
11	0.581	0.530	0.453	0.387
12	0.548	0.505	0.430	0.365
13	0.527	0.481	0.410	0.347
14	0.503	0.458	0.391	0.331
15	0.482	0.439	0.375	0.317
16	0.466	0.423	0.361	0.305
17	0.450	0.410	0.348	0.294
18	0.434	0.395	0.336	0.284
19	0.421	0.382	0.326	0.275
20	0.410	0.372	0.317	0.267
25	0.362	0.328	0.278	0.235
30	0.328	0.297	0.251	0.211

Source: Adapted from S. Maghsoodloo (1975), Estimates of the quantiles of Kendall's partial rank correlation coefficient and additional quantile estimates, *Journal of Statistical Computation and Simulation*, **4**, 155–164, and S. Maghsoodloo, and L. L. Pallos (1981), Asymptotic behavior of Kendall's partial rank correlation coefficient and additional quantile estimates, *Journal of Statistical Computation and Simulation*, **13**, 41–48, with permission.

Table Q Page's *L* Statistic

Each table entry for n treatments and k blocks is the value of L such that its right-tail probability is less than or equal to 0.001 for the upper number, 0.01 for the middle number, and 0.05 for the lower number.

	n					
k	3	4	5	6	7	8
			109	178	269	388
2		60	106	173	261	376
	28	58	103	166	252	362
		89	160	260	394	567
3	42	87	155	252	382	549
	41	84	150	244	370	532
	56	117	210	341	516	743
4	55	114	204	331	501	722
	54	111	197	321	487	701
	70	145	259	420	637	917
5	68	141	251	409	620	893
	66	137	244	397	603	869
	83	172	307	499	757	1090
6	81	167	299	486	737	1063
	79	163	291	474	719	1037
	96	198	355	577	876	1262
7	93	193	346	563	855	1232
	91	189	338	550	835	1204
	109	225	403	655	994	1433
8	106	220	383	640	972	1401
	104	214	384	625	950	1371
	121	252	451	733	1113	1603
9	119	246	441	717	1088	1569
	116	240	431	701	1065	1537
	134	278	499	811	1230	1773
10	131	272	487	793	1205	1736
	128	266	477	777	1180	1703
	147	305	546	888	1348	1943
11	144	298	534	869	1321	1905
	141	292	523	852	1295	1868
	160	331	593	965	1465	2112
12	156	324	581	946	1437	2072
	153	317	570	928	1410	2035

Source: Adapted from E. P. Page (1963), Ordered hypotheses for multiple treatments: A significance test for linear ranks, *Journal of the American Statistical Association*, **58**, 216–230, with permission.

Table R Critical Values and Associated Probabilities for the Jonckheere-Terpstra Test

Each entry is the critical value $B_\alpha = B(\alpha, k, n_1, n_2, \ldots, n_k)$ of a Jonckheere-Terpstra statistic B for given $\alpha, k, n, n_2, \ldots, n_k$, such that $P(B \geqslant B_\alpha | H_0) \leqslant \alpha$. The actual right-tail probability is equal to the value given in parentheses.

			$k=3$						
n_1	n_2	n_3	$\alpha = 0.5$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
2	2	2	7 (.42222)	9 (.16667)	10 (.08889)	11 (.03333)	12 (.01111)	—	—
2	2	3	9 (.43810)	12 (.13810)	13 (.07619)	14 (.03810)	15 (.01429)	16 (.00476)	16 (.00476)
2	2	4	11 (.44762)	14 (.18095)	16 (.07143)	17 (.03810)	18 (.01905)	19 (.00714)	20 (.00238)
2	2	5	13 (.45503)	17 (.15344)	19 (.06614)	20 (.03968)	21 (.02116)	23 (.00397)	23 (.00397)
2	2	6	15 (.46032)	19 (.18492)	21 (.09444)	23 (.03968)	24 (.02381)	26 (.00635)	27 (.00238)
2	2	7	17 (.46465)	22 (.16364)	24 (.08788)	26 (.04040)	28 (.01515)	29 (.00808)	30 (.00404)
2	2	8	19 (.46801)	24 (.18855)	27 (.08215)	29 (.04040)	31 (.01684)	33 (.00539)	34 (.00269)
2	3	3	12 (.40000)	15 (.15179)	16 (.09643)	18 (.03036)	19 (.01429)	20 (.00536)	21 (.00179)
2	3	4	14 (.45714)	18 (.16190)	20 (.07381)	21 (.04524)	23 (.01349)	24 (.00635)	25 (.00138)
2	3	5	17 (.42500)	21 (.16944)	23 (.08770)	25 (.03810)	26 (.02302)	28 (.00675)	29 (.00317)
2	3	6	19 (.46645)	24 (.17554)	26 (.09957)	28 (.04957)	30 (.02100)	32 (.00714)	33 (.00368)
2	3	7	22 (.44003)	27 (.18030)	30 (.08232)	32 (.04268)	34 (.01032)	36 (.00732)	37 (.00368)
2	3	8	24 (.47273)	30 (.18430)	33 (.09192)	36 (.03768)	38 (.01810)	40 (.00754)	41 (.00451)
2	4	4	17 (.46254)	21 (.19810)	24 (.07556)	26 (.03206)	27 (.01905)	29 (.00540)	30 (.00254)
2	4	5	20 (.46739)	25 (.18095)	28 (.07662)	30 (.03680)	31 (.02395)	33 (.00880)	34 (.00491)
2	4	6	23 (.47071)	29 (.16797)	32 (.07742)	34 (.04076)	36 (.01898)	38 (.00758)	39 (.00440)
2	4	7	26 (.47366)	32 (.19305)	36 (.07797)	38 (.04406)	40 (.02261)	43 (.00660)	44 (.00408)
2	4	8	29 (.47590)	36 (.18077)	39 (.09879)	42 (.04686)	45 (.01863)	47 (.00892)	49 (.00377)
2	5	5	24 (.44228)	29 (.19000)	32 (.09157)	35 (.03565)	36 (.02453)	39 (.00643)	40 (.00373)
2	5	6	27 (.47403)	33 (.19708)	37 (.08178)	39 (.04715)	41 (.02486)	44 (.00777)	45 (.00491)
2	5	7	31 (.45303)	38 (.17057)	41 (.09355)	44 (.04477)	47 (.01820)	49 (.00894)	51 (.00393)
2	5	8	34 (.47849)	42 (.17764)	46 (.08500)	49 (.04283)	52 (.01885)	54 (.00996)	56 (.00482)
2	6	6	31 (.47662)	38 (.18816)	42 (.08528)	45 (.04027)	47 (.02235)	50 (.00786)	52 (.00343)

2	6	7	35	(.47875)	43	(.18087)	47	(.08803)	50	(.04521)	53	(.02040)	56	(.00789)	58	(.00376)
2	6	8	39	(.48051)	48	(.17491)	52	(.09031)	55	(.04953)	58	(.02449)	62	(.00788)	64	(.00404)
2	7	7	40	(.46130)	48	(.18948)	53	(.08358)	56	(.04543)	59	(.02225)	62	(.00964)	65	(.00360)
2	7	8	44	(.48219)	53	(.19675)	58	(.09468)	62	(.04555)	65	(.02381)	69	(.00857)	71	(.00474)
2	8	8	49	(.48359)	59	(.19248)	64	(.09833)	69	(.04231)	72	(.02319)	76	(.00913)	79	(.00404)
3	3	3	15	(.41548)	18	(.19405)	20	(.09464)	22	(.03690)	23	(.02083)	25	(.00476)	25	(.00476)
3	3	4	18	(.42667)	22	(.17500)	24	(.09310)	26	(.04214)	28	(.01548)	29	(.00857)	30	(.00429)
3	3	5	21	(.43528)	26	(.16147)	28	(.09177)	30	(.04621)	32	(.02002)	34	(.00714)	35	(.00390)
3	3	6	24	(.44210)	29	(.18912)	32	(.09075)	34	(.04946)	36	(.02408)	39	(.00622)	40	(.00357)
3	3	7	27	(.44761)	33	(.17619)	36	(.08989)	39	(.03849)	41	(.01941)	43	(.00868)	45	(.00335)
3	3	8	30	(.45216)	36	(.19843)	40	(.08914)	43	(.04144)	45	(.02259)	48	(.00759)	49	(.00498)
3	4	4	21	(.46779)	26	(.18528)	29	(.08043)	31	(.03974)	33	(.01688)	35	(.00589)	36	(.00320)
3	4	5	25	(.44304)	30	(.19325)	33	(.09481)	36	(.03791)	38	(.01789)	40	(.00732)	41	(.00440)
3	4	6	28	(.47434)	34	(.19973)	38	(.08432)	40	(.04923)	43	(.01865)	45	(.00856)	47	(.00343)
3	4	7	32	(.45344)	39	(.17279)	42	(.09566)	45	(.04644)	48	(.01926)	50	(.00963)	52	(.00435)
3	4	8	35	(.47863)	43	(.17947)	47	(.08672)	50	(.04419)	53	(.01974)	56	(.00752	58	(.00354)
3	5	5	29	(.44913)	35	(.18365)	38	(.09706)	41	(.04382)	43	(.02324)	46	(.00740)	47	(.00475)
3	5	6	33	(.45405)	40	(.17607)	43	(.09882)	46	(.04890)	49	(.02085)	52	(.00741)	54	(.00328)
3	5	7	37	(.45809)	44	(.19851)	49	(.08220)	52	(.04211)	55	(.01903)	58	(.00740)	60	(.00356)
3	5	8	41	(.46147)	49	(.19057)	54	(.08461)	57	(.04627)	60	(.02284)	64	(.00737)	66	(.00380)
3	6	6	37	(.47913)	45	(.18533)	49	(.09226)	52	(.04855)	55	(.02267)	58	(.00919)	60	(.00459)
3	6	7	42	(.46187)	50	(.19315)	55	(.08704)	58	(.04821)	61	(.02420)	65	(.00807)	67	(.00425)

(Continued)

Table R (*Continued*)

			k = 3													
n_1	n_2	n_3	$\alpha = 0.5$		$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.01$		$\alpha = 0.005$	
3	6	8	46	(.48241)	55	(.19983)	60	(.09770)	64	(.04788)	68	(.02027)	71	(.00950)	74	(.00397)
3	7	7	47	(.46502)	56	(.18868)	61	(.09112)	65	(.04421)	68	(.02337)	72	(.00861)	74	(.00486)
3	7	8	52	(.46769)	62	(.18490)	67	(.09460)	71	(.04917)	75	(.02265)	79	(.00907)	82	(.00410)
3	8	8	57	(.48503)	68	(.19289)	74	(.09197)	79	(.04251)	82	(.02477)	87	(.00869)	90	(.00418)
4	4	4	25	(.47160)	31	(.17558)	34	(.08439)	36	(.04632)	38	(.02286)	40	(.00993)	42	(.00367)
4	4	5	29	(.47465)	36	(.16825)	39	(.08738)	42	(.03873)	44	(.02027)	46	(.00959)	48	(.00402)
4	4	6	33	(.47708)	40	(.19294)	44	(.08984)	47	(.04376)	49	(.02497)	52	(.00931)	54	(.00429)
4	4	7	37	(.47909)	45	(.18488)	49	(.09184)	52	(.04822)	55	(.02244)	58	(.00906)	60	(.00450)
4	4	8	41	(.48077)	50	(.17830)	54	(.09353)	58	(.04204)	61	(.02049)	64	(.00885)	66	(.00468)
4	5	5	34	(.45453)	41	(.17872)	45	(.08177)	48	(.03928)	50	(.02220)	53	(.00815)	55	(.00371)
4	5	6	38	(.47932)	46	(.18750)	50	(.09435)	54	(.03970)	56	(.02382)	59	(.00987)	62	(.00347)
4	5	7	43	(.46215)	51	(.19494)	56	(.08875)	59	(.04959)	63	(.01963)	66	(.00858)	68	(.00459)
4	5	8	47	(.48252)	57	(.17722)	61	(.09919)	65	(.04905)	69	(.02102)	72	(.00998)	75	(.00425)
4	6	6	43	(.48114)	52	(.18307)	56	(.09810)	60	(.04546)	63	(.02287)	67	(.00769)	69	(.00408)
4	6	7	48	(.48267)	58	(.17938)	63	(.08619)	67	(.04174)	70	(.02208)	74	(.00818)	76	(.00463)
4	6	8	53	(.48397)	63	(.19820)	69	(.08972)	73	(.04561)	77	(.02141)	81	(.00859)	84	(.00390)
4	7	7	54	(.46809)	64	(.18800)	69	(.09759)	74	(.04318)	77	(.02426)	82	(.00783)	84	(.00465)
4	7	8	59	(.48519)	70	(.19552)	76	(.09450)	81	(.04441)	85	(.02170)	89	(.00946)	92	(.00466)
4	8	8	65	(.48624)	77	(.19320)	83	(.09869)	88	(.04966)	93	(.02191)	98	(.00831)	101	(.00428)
5	5	5	39	(.45888)	47	(.17478)	51	(.08666)	54	(.04558)	57	(.02136)	60	(.00873)	62	(.00440)
5	5	6	44	(.46248)	52	(.19706)	57	(.09078)	61	(.04151)	64	(.02066)	67	(.00921)	70	(.00360)
5	5	7	49	(.46569)	58	(.19200)	63	(.09430)	67	(.04665)	71	(.02008)	74	(.00960)	77	(.00413)
5	5	8	54	(.46806)	64	(.18774)	69	(.09734)	74	(.04300)	77	(.02413)	81	(.00992)	84	(.00461)
5	6	6	49	(.48280)	59	(.18118)	64	(.08787)	68	(.04301)	71	(.02299)	75	(.00868)	77	(.00498)
5	6	7	55	(.46829)	65	(.18952)	70	(.09906)	75	(.04427)	79	(.02042)	83	(.00824)	85	(.00494)
5	6	8	60	(.48527)	71	(.19681)	77	(.09575)	82	(.04535)	86	(.02235)	90	(.00985)	93	(.00490)
5	7	7	61	(.47066)	72	(.18741)	78	(.09019)	82	(.04981)	86	(.02499)	91	(.00904)	94	(.00447)
5	7	8	67	(.47271)	79	(.18559)	85	(.09438)	90	(.04739)	94	(.02489)	99	(.00077)	103	(.00410)

5	8	8	73	(.48727)	86	(.19344)	93	(.09319)	98	(.04917)	103	(.02322)	108	(.00968)	112	(.00434)
6	6	6	55	(.48418)	66	(.17959)	71	(.09285)	75	(.04897)	79	(.02306)	83	(.00954)	86	(.00447)
6	6	7	61	(.48536)	72	(.19831)	78	(.09721)	83	(.04645)	87	(.02311	92	(.00827)	95	(.00406)
6	6	8	67	(.48638)	79	(.19561)	86	(.08914)	91	(.04436)	95	(.02313)	100	(.00899)	103	(.00471)
6	7	7	68	(.47285)	80	(.18689)	86	(.09563)	91	(.04835)	96	(.02153)	101	(.00829)	104	(.00432)
6	7	8	74	(.48733)	87	(.19456)	94	(.09426)	100	(.04351)	104	(.02380)	110	(.00829)	113	(.00454)
6	8	8	81	(.48816)	95	(.19364)	102	(.09885)	108	(.04873)	113	(.02437)	119	(.00923)	123	(.00439)
7	7	7	75	(.47473)	88	(.18643)	95	(.08944)	100	(.04711)	105	(.02225)	110	(.00929)	114	(.00418)
7	7	8	82	(.47637)	96	(.18602)	103	(.09414)	109	(.04605)	114	(.02288)	120	(.00859)	123	(.00494)
7	8	8	89	(.48892)	104	(.19380)	112	(.09402)	118	(.04834)	124	(.02209)	130	(.00885)	134	(.00443)
8	8	8	97	(.48960)	113	(.19393)	121	(.09891)	128	(.04798)	34	(.02310)	140	(.00992)	145	(.00445)

				$k=4$													
n_1	n_2	n_3	n_4	$\alpha = 0.5$		$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.01$		$\alpha = 0.005$	
2	2	2	2	13	(.45079)	16	(.19286)	18	(.08294)	19	(.04841)	21	(.01230)	22	(.00516)	23	(.00159)
3	3	3	3	28	(.47240)	34	(.18229)	37	(.09067)	40	(.03744)	42	(.01834)	44	(.00797)	45	(.00498)
4	4	4	4	49	(.48174)	58	(.19096)	63	(.08950)	67	(.04198)	70	(.02150)	73	(.00998)	76	(.00414)
5	5	5	5	76	(.48679)	89	(.18455)	95	(.09621)	100	(.04983)	105	(.02296)	110	(.00928)	114	(.00404)
6	6	6	6	109	(.48987)	126	(.18631)	134	(.09607)	141	(.04743)	147	(.02336)	154	(.00894)	158	(.00481)

(Continued)

Table R (*Continued*)

n_1	n_2	n_3	n_4	n_5	$\alpha = 0.5$		$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.01$		$\alpha = 0.005$	
					$k=5$													
2	2	2	2	2	21	(.46466)	26	(.16246)	28	(.08779)	30	(.04116)	32	(.01623)	33	(.00939)	35	(.00257)
3	3	3	3	3	46	(.48020)	54	(.19822)	59	(.08738)	62	(.04752)	65	(.02335)	69	(.00755)	71	(.00392)
4	4	4	4	4	81	(.48695)	94	(.18756)	100	(.09910)	106	(.04523)	110	(.02450)	116	(.00839)	119	(.00455)
5	5	5	5	5	126	(.49058)	144	(.19032)	153	(.09542)	160	(.04970)	167	(.02318)	174	(.00958)	179	(.00470)
6	6	6	6	6	181	(.49279)	204	(.19719)	216	(.09842)	226	(.04884)	235	(.02292)	244	(.00969)	251	(.00456)

n_1	n_2	n_3	n_4	n_5	n_6	$\alpha = 0.5$		$\alpha = 0.2$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.01$		$\alpha = 0.005$	
						$k=6$													
2	2	2	2	2	2	31	(.47293)	37	(.18713)	40	(.09533)	43	(.04083)	45	(.02071)	47	(.00944)	49	(.00379)
3	3	3	3	3	3	69	(.46981)	80	(.18058)	85	(.09686)	90	(.04524)	94	(.02201)	98	(.00958)	101	(.00473)
4	4	4	4	4	4	121	(.49006)	138	(.19092)	147	(.09181)	154	(.04567)	160	(.02274)	167	(.00888)	171	(.00486)
5	5	5	5	5	5	189	(.48567)	212	(.19377)	224	(.09679)	234	(.04775)	242	(.02477)	252	(.00964)	259	(.00456)
6	6	6	6	6	6	271	(.49452)	302	(.19304)	317	(.09982)	330	(.04490)	342	(.02361)	354	(.00999)	363	(.00485)

Because of the symmetry of the distribution of B under H_0 critical values and associated probabilities for all possible sample size combinations from 2 to 8 from 3 populations can be obtained from the table. For example, if $n_1 = 4$, $n_2 = 5$, and $n_3 = 2$, the required values are given by the table entry for $n_1 = 2$, $n_2 = 4$, and $n_3 = 5$.

Source: Adapted from Tables 1 and 2 of R. E. Odeh (1971), On Jonckheere's k-sample test against ordered alternatives, *Technometrices*, **13**, 912–918, with permission.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table S Rank von Neumann Statistic

Each table entry for $n \leqslant 10$ is the exact left-tail or right-tail P value of the corresponding listed value of NM. Only those values of NM that are close to the typical values of $\alpha = 0.005$, 0.01, 0.025, 0.05 and 0.10 are included. The table entries for $n > 10$ are the left-tail critical values of RVN for the same typical α values. Since these entries are based on a beta approximation which is symmetric about 2, corresponding right-tail critical values are easily found. For example if $n = 40$, $\alpha = 0.005$, the left-tail critical value of RVN is 1.22 and hence the right-tail critical value is 2.78.

	P values for selected values of NM			
n	NM	*Left-tail P*	NM	*Right-tail P*
4	3	0.0833	17	0.0833
	6	0.2500	14	0.2500
5	4	0.0167	35	0.0333
	7	0.0500	33	0.0667
	10	0.1333	30	0.1333
6	5	0.0028	65	0.0028
	8	0.0083	63	0.0083
	11	0.0250	62	0.0139
	14	0.0472	60	0.0194
	16	0.0750	59	0.0306
	17	0.0806	56	0.0361
	19	0.1306	55	0.0694
			52	0.0972
			51	0.1139
7	14	0.0048	101	0.0040
	15	0.0079	100	0.0056
	17	0.0119	98	0.0087
	18	0.0151	97	0.0103
	20	0.0262	93	0.0206
	24	0.0444	92	0.0254
	25	0.0563	88	0.0464
	31	0.0988	87	0.0536
	32	0.1155	81	0.0988
			80	0.1115
8	23	0.0049	149	0.0043
	24	0.0073	148	0.0052
	26	0.0095	144	0.0084
	27	0.0111	143	0.0105
	32	0.0221	136	0.0249
	33	0.0264	135	0.0286
	39	0.0481	129	0.0481
	40	0.0529	128	0.0530
	48	0.0978	120	0.0997

(*Continued*)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table S (*Continued*)

n	P values for selected values of NM			
	NM	*Left-tail P*	NM	*Right-tail P*
	49	0.1049	119	0.1074
9	34	0.0045	208	0.0046
	35	0.0055	207	0.0053
	40	0.0096	202	0.0091
	41	0.0109	201	0.0104
	49	0.0236	191	0.0245
	50	0.0255	190	0.0262
	59	0.0486	181	0.0499
	60	0.0516	180	0.0528
	71	0.0961	169	0.0978
	72	0.1010	168	0.1030
10	51	0.0050	282	0.0046
	59	0.0100	281	0.0051
	72	0.0242	273	0.0097
	73	0.0260	272	0.0103
	85	0.0493	259	0.0240
	86	0.0517	258	0.0252
	101	0.0985	246	0.0475
	102	0.1017	245	0.0504
			229	0.0990
			228	0.1023

n	Left-tail critical values of RVN				
	0.005	*0.010*	*0.025*	*0.050*	*0.100*
10	0.62	0.72	0.89	1.04	1.23
11	0.67	0.77	0.93	1.08	1.26
12	0.71	0.81	0.96	1.11	1.29
13	0.74	0.84	1.00	1.14	1.32
14	0.78	0.87	1.03	1.17	1.34
15	0.81	0.90	1.05	1.19	1.36
16	0.84	0.93	1.08	1.21	1.38
17	0.87	0.96	1.10	1.24	1.40
18	0.89	0.98	1.13	1.26	1.41
19	0.92	1.01	1.15	1.27	1.43
20	0.94	1.03	1.17	1.29	1.44

(*Continued*)

Table S (*Continued*)

n	*Left-tail critical values of* RVN 0.005	0.010	0.025	0.050	0.100
21	0.96	1.05	1.18	1.31	1.45
22	0.98	1.07	1.20	1.32	1.46
23	1.00	1.09	1.22	1.33	1.48
24	1.02	1.10	1.23	1.35	1.49
25	1.04	1.12	1.25	1.36	1.50
26	1.05	1.13	1.26	1.37	1.51
27	1.07	1.15	1.27	1.38	1.51
28	1.08	1.16	1.28	1.39	1.52
29	1.10	1.18	1.30	1.40	1.53
30	1.11	1.19	1.31	1.41	1.54
32	1.13	1.21	1.33	1.43	1.55
34	1.16	1.23	1.35	1.45	1.57
36	1.18	1.25	1.36	1.46	1.58
38	1.20	1.27	1.38	1.48	1.59
40	1.22	1.29	1.39	1.49	1.60
42	1.24	1.30	1.41	1.50	1.61
44	1.25	1.32	1.42	1.51	1.62
46	1.27	1.33	1.43	1.52	1.63
48	1.28	1.35	1.45	1.53	1.63
50	1.29	1.36	1.46	1.54	1.64
55	1.33	1.39	1.48	1.56	1.66
60	1.35	1.41	1.50	1.58	1.67
65	1.38	1.43	1.52	1.60	1.68
70	1.40	1.45	1.54	1.61	1.70
75	1.42	1.47	1.55	1.62	1.71
80	1.44	1.49	1.57	1.64	1.71
85	1.45	1.50	1.58	1.65	1.72
90	1.47	1.52	1.59	1.66	1.73
95	1.48	1.53	1.60	1.66	1.74
100	1.49	1.54	1.61	1.67	1.74
100[a]	1.48	1.53	1.61	1.67	1.74
100[b]	1.49	1.54	1.61	1.67	1.74

[a]Using the $N(2, 4/n)$ approximation.
[b]Using the $N[2, 20/(5n+7)]$ approximation.
Source: Adapted from R. Bartels (1982), The rank version of von Neumann's ratio test for randomness, *Journal of the American Statistical Association*, **77**, 40–46, with permission.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table T Lilliefors's Test for Exponential Distribution Critical Values
Table entries for any sample size N are the values of a Lilliefors's random variable with right-tail probability as given in the top row.

Sample Size	*Significance Level*			
N	*0.100*	*0.050*	*0.010*	*0.001*
4	.444	.483	.556	.626
5	.405	.443	.514	.585
6	.374	.410	.477	.551
7	.347	.381	.444	.509
8	.327	.359	.421	.502
9	.310	.339	.399	.460
10	.296	.325	.379	.444
11	.284	.312	.366	.433
12	.271	.299	.350	.412
14	.252	.277	.325	.388
16	.237	.261	.311	.366
18	.224	.247	.293	.328
20	.213	.234	.279	.329
25	.192	.211	.251	.296
30	.176	.193	.229	.270
40	.153	.168	.201	.241
50	.137	.150	.179	.214
60	.125	.138	.164	.193
75	.113	.124	.146	.173
100	.098	.108	.127	.150
Over 100	$.980/\sqrt{N}$	$1.077/\sqrt{N}$	$1.274/\sqrt{N}$	$1.501/\sqrt{N}$

Source: Adapted from R. L. Edgeman and R. C. Scott (1987), Lilliefors's tests for transformed variables, *Brazilian Journal of Probability and Statistics*, 1, 101–112, with permission.

Answers to Selected Problems

2.6 $Y = 4X - 2X^2$

2.7 $\left(\frac{7-x}{6}\right)^5 - \left(\frac{6-x}{6}\right)^5, \quad x = 1, 2, \ldots, 6$

2.8 $X_{(1)} - \ln(20/3)$

2.10 $(a)\ 1 - (0.9)^{10} \qquad (b)\ 1 - (0.5)^{1/10}$

2.11 $(a)\ 11/6 \qquad (b)\ 3/(2\sqrt{\pi})$

2.12 $8u^2(3 - 4u), \quad 0 < u < 1/2$
$32u^3 - 72u^2 + 48u - 8, \quad 1/2 < u < 1$

2.13 $(a)\ 1/2,\ 1/4(n+2)$
$(b)\ 1/2,\ n/4(n+1)^2$

2.14 $(n-1)(e^8 - 1)^2/2$

2.15 $4(n-1)e^{4u}(e^{4u} - 1)$

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

2.16 $n(2u)^{n-1}, \quad 0 < u < 1/2$
$n[2(1-u)]^{n-1}, \quad 1/2 < u < 1$

2.18 (*a*) $\mu,\ \pi\sigma^2/2(2m+3)$ (*b*) $0.2877,\ 0.016$

2.23 0.66

2.24 0.50

2.25 0.05

2.26 $n(n-1); 2(n-1)/(n+1)^2(n+2)$

3.15 (*a*) $P = 0.0012$
(*b*) No, too many zeros

3.16 (*a*) $R = 2,\ P = 0.095$
(*b*) $R = 2,\ P = 0.025$

3.17 (*a*) $R = 6,\ P = 0.069;\ R = 11,\ P = 0.3770$
(*b*) $R = 4,\ P = 0.054;\ R = 10,\ P = 0.452$
(*c*) $R = 5,\ P = 0.024;\ R = 12,\ P = 0.3850$

3.18 $R = 6,\ P > 0.5;\ R = 5,\ P > 0.7573$

4.1 $Q = 3.1526,\ 0.25 < P < 0.50$

4.2 $Q = 7.242,\ 0.10 < P < 0.25$

4.12 $n = 1,063$

4.18 $D = 0.2934,\ 0.10 < P < 0.20$

4.20 (*a*) $D = 0.3117,\ 0.05 < P < 0.10$
(*b*) $D = 0.1994,\ P > 0.20$
(*c*) (i) 28 (ii) 47

4.21 (*a*) $Q = 76.89,\ \text{df} = 9,\ P < 0.001$ or $Q = 61.13$,
$\text{df} = 7,\ P < 0.001$

4.25 $Q = 0.27,\ \text{df} = 1,\ P > 0.50$

4.27 $Q = 35.54,\ P < 0.001$

4.28 *K-S*

4.30 (*a*) $D = 0.1813,\ P > 0.20$
(*b*) $D = 0.1948,\ P > 0.20$

4.34 $D = 0.400,\ 0.05 < P < 0.10$

5.2 (*a*) $[(N-1)/(N+1)]^{1/2}$ (*b*) $1/(2\sqrt{\pi})$
(*c*) $[3(N-1)/4(N+1)]^{1/2}$

5.4 (i) (a) Reject, $K \geqslant 6$ (b) 0.063
(c) $K = 4$, do not reject H_0 (d) 0.070
(e) $-4 \leqslant M_D \leqslant 16$

(ii) (a) Do not reject H_0, $T^- = 6$ (b) 0.08
(c) Do not reject H_0, $T^+ = 16.5$ (d) 0.039
(e) $-1.5 \leqslant M_D \leqslant 7$

5.10 $T^+ = 53,\ P = 0.003;\ K = 9,\ P = 0.0107$

5.12 $K = 12,\ P = 0.0017$

5.13 $K = 13,\ P = 0.0835$

5.14 $249 \leqslant M_D \leqslant 1157,\ \gamma = 0.9376$
$273 \leqslant M_D \leqslant 779,\ \gamma = 0.876$

5.15 $K = 4,\ P = 0.1875$

5.16 (a) $-8 \leqslant \mathrm{M} \leqslant 10,\ \gamma = 0.961$
(b) $-7 \leqslant \mathrm{M} \leqslant 7.5,\ \gamma = 0.96$

5.20 (a) $T^+ = 48,\ P = 0.019$
(b) $0.5 \leqslant M_D \leqslant 6,\ \gamma = 0.916$
(c) $K = 7,\ P = 0.1719$
(d) $-2 \leqslant M_D \leqslant 6,\ \gamma = 0.9786$

5.21 $P = 0.0202$

5.22 (a) $K = 4,\ P = 0.3438$
(b) $-4 \leqslant M_D \leqslant 13,\ \gamma = 0.9688$
(c) $T^+ = 18,\ P = 0.078$
(d) $-3 \leqslant M_D \leqslant 9.5,\ \gamma = 0.906$

5.23 $P = 0.0176$

5.29 (a) 15/16 (b) $5^2 3^3/4^5$ (c) $(0.8)^4$

6.1 0.75

6.6 (i) (a) $U = 6$, do not reject H_0 (b) 130/12,870
(c) $-27 \leqslant \theta \leqslant 80$

(ii) (a) $U = 12.5$, reject H_0 (b) $\alpha = 0.082$
(c) $12 \leqslant M_Y - M_X \leqslant 65$

6.9 $mnD = 54,\ P = 0.0015$
$R = 4,\ P = 0.010$

6.14 $P = 0.07,\ 3 \leqslant \theta \leqslant 9,\ \gamma = 0.857$

8.9 $-7 \leqslant M_X - M_Y \leqslant 19,\ \gamma = 0.97$
$-6 \leqslant M_X - M_Y \leqslant 15,\ \gamma = 0.948$

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

8.10 $W_N = 60,\ P = 0.086$

8.13 $W_N = 54,\ P = 0.866$

8.14 $(a)\ u = 3,\ \gamma = 0.9346$
$(b)\ u = 14,\ \gamma = 0.916$
$(c)\ u = 6,\ \gamma = 0.918$

8.15 $(a)\ W_N = 14,\ P = 0.452$
$(b)\ -17 \leqslant M_D \leqslant 24,\ \gamma = 0.904$

9.11 $P = 0.042$

9.13 $(a)\ 0.228 < P < 0.267$
$(b)\ 0.159 < P < 0.191$
(c) K-S or chi square

10.12 $Q = 3.24$

10.13 $H = 18.91$

10.14 $S = 43.5,\ 0.77 < P < 0.094$

10.16 $(a)\ Q = 43.25,\ \text{df} = 2,\ P < 0.001$

10.17 $H = 16.7,\ H_c = 17.96,\ df = 2,\ P < 0.001$

10.22 $H = 10.5,\ 0.001 < P < 0.01$

11.1 (i) $(a)\ T = 0.5$ (b) Do not reject H_0
$(c)\ 0.43 \leqslant \tau \leqslant 0.57$
(ii) $(a)\ R = 0.69$ (b) Do not reject H_0

11.3 $(a)\ T = 2(mn - 2u)/N(N-1)$ $(b)\ T = (2u - mn)/\binom{N}{2}$

11.5 $R = 0.25,\ P = 0.26$
$R_c = 0.244,\ P \approx 0.268$
$T = 0.17,\ P = 0.306$
$T_c = 0.17,\ P \approx 0.3$

11.6 $(a)\ T^+ = 57,\ z = 0.25,\ P = 0.4013$
$(b)\ T = 0.648,\ P < 0.005$
$R = 0.8038,\ P < 0.001$

11.7 $R = 0.661,\ 0.029 < P < 0.033$
$T = 0.479,\ 0.038 < P < 0.06$
$\text{NM} = 24.75,\ P < 0.0045$
R (runs up and down) $= 3,\ P = 0.0257$

11.14 $(a)\ R = 0.7143$
$(b)\ P = 0.068$

(*c*) $T = 0.4667$
(*d*) $P = 0.136$

11.15 $R = 0.6363,\ P = 0.027$
$T = 0.51,\ P = 0.023$

11.18 (*a*) $T = 0.6687$
(*b*) $R = 0.7793$

12.1 $S = 37,\ P = 0.033$

12.2 (*a*) $S = 312,\ W = 0.825$

12.4 $T_{XY} = 7/15,\ T_{XZ} = 5/15$
$T_{YZ} = 5/15,\ T_{XY.Z} = 2/5$

12.5 (a) $R = 0.977$

12.6 $T_{XZ} = 0.80,\ T_{YZ} = -0.9487$
$T_{XY} = -0.7379,\ T_{XY.Z} = 0.1110$

12.7 $T_{12} = -0.7143,\ T_{13} = -0.5714$
$T_{23} = 0.5714,\ T_{23.1} = 0.2842$
$P > 0.05$

12.8 $Q = 4.814,\ \text{df} = 5,\ P > 0.30$
$Q_C = 4.97,\ P > 0.30$

12.9 $Q = 19.14,\ \text{df} = 6,\ 0.001 < P < 0.005$

12.13 (*a*) $W = 0.80,\ P < 0.02$ (*b*) IADGBHCJFE

13.1 (*a*) $1/2\sigma\sqrt{\pi}$ (*b*) $3N\lambda_N(1-\lambda_N)$ (*c*) $3N\lambda_N(1-\lambda_N)\lambda^2/4$

13.2 (*b*) $1/12$ (*c*) σ^2

13.5 1

14.2 $Q = 8.94,\ \text{df} = 2,\ 0.01 < P < 0.05$

14.3 $Z_c = 1.94,\ P = 0.0262$

14.4 $Z_c = 0.80,\ P = 0.2119$

14.5 Exact $P = 0.2619$

14.6 $Z_c = 1.643,\ P = 0.0505$

14.7 $Z_c = 3.82,\ P < 0.001$

14.8 $Q = 0.5759,\ \text{df} = 2,\ 0.50 < P < 0.70$

14.9 (*a*) $Q = 13.83,\ \text{df} = 4,\ 0.005 < P < 0.01$
(*b*) $C = 0.35,\ \phi = 0.37$

(c) $T = -0.16,\ Z = -2.39,\ P = 0.0091$
(d) $\gamma = -0.2366$

14.10 $Z_c = 0.62,\ P = 0.2676$

14.11 Exact $P = 0.0835$

14.12 Exact $P = 0.1133$

14.13 $Q = 261.27,\ \text{df} = 12,\ P < 0.001$

14.14 $Z = 2.01,\ P = 0.0222$

References

Abramowitz, M., and I. A. Stegun (1972), *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematics Series **55**, Washington, D.C., 10th Printing.

Alexander, D. A., and D. Quade (1968), On the Kruskal-Wallis three sample H-statistic, *University of North Carolina, Department of Biostatistics, Institute of Statistics Mimeo Series* 602, Chapel Hill, NC.

Andrews, F. C. (1954), Asymptotic behavior of some rank tests for analysis of variance, *Annals of Mathematical Statistics*, **25**, 724–736.

Andrews, J. C. (1989), The dimensionality of beliefs toward advertising in general, *Journal of Advertising*, **18**, 26–35.

Ansari, A. R., and R. A. Bradley (1960), Rank sum tests for dispersion, *Annals of Mathematical Statistics*, **31**, 1174–1189.

Arnold, H. J. (1965), Small sample power for the one sample Wilcoxon test for non-normal shift alternatives, *Annals of Mathematical Statistics*, **36**, 1767–1778.

Auble, J. D. (1953), Extended tables for the Mann-Whitney statistic, *Bulletin of the Institute for Educational Research*, **1**, i–iii, and 1–39.

Aubuchon, J. C., and T. P. Hettmansperger (1984), A note on the estimation of the integral of f^2x, *Journal of Statistical Planning and Inference*, **9**, 321–331.

Bahadur, R. R. (1960*a*), Asymptotic efficiency of tests and estimators, *Sankhyā*, **B20**, 229–252.

Bahadur, R. R. (1960*b*), Stochastic comparison of test, *Annals of Mathematical Statistics*, **31**, 276–295.

Bahadur, R. R. (1967), Rates of convergence of estimates and test statistics, *Annals of Mathematical Statistics*, **38**, 303–324.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972), *Statistical Inference Under Order Restrictions*, John Wiley & Sons, New York.

Bartels, R. (1982), The rank version of von Neumann's ratio test for randomness, *Journal of the American Statistical Association*, **77**, 40–46.

Bateman, G. (1948), On the power function of the longest run as a test for randomness in a sequence of alternatives, *Biometrika*, **35**, 97–112.

Berenson, M. L. (1982), A comparison of several k sample tests for ordered alternatives in completely randomized designs, *Psychometrika*, **47**, 265–280.

Bergmann, R., J. Ludbrook, and W. P. J. Spooren (2000), Different outcomes from the Mann-Whitney test from different statistical packages, *The American Statistician*, **34**, 72–77.

Best, D. J. (1973), Extended tables for Kendall's tau, *Biometrika*, **60**, 429–430.

Best, D. J. (1974), Tables for Kendall's tau and an examination of the normal approximation, *Technical Paper No. 39*, Division of Mathematical Statistics, Commonwealth Scientific and Industrial Research Organization Australia.

Best, D. J., and P. G. Gipps (1974), The upper tail probabilities of Kendall's tau, *Applied Statistics*, **23**, 98–100.

Birnbaum, Z. W. (1952), Numercial tabulation of the distribution of Kolmogorov's statistic for finite sample size, *Journal of the American Statistical Association*, **47**, 425–441.

Birnbaum, Z. W., and F. H. Tingey (1951), One-sided confidence contours for probability distribution functions, *Annals of Mathematical Statistics*, **22**, 592–596.

Blair, R. C., and J. J. Higgins (1980), A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various nonnormal distributions, *Journal of Educational Statistics*, **5**, 309–335.

Blom, G. (1958), *Statistical estimates and transformed Beta-variables*, John Wiley & Sons, New York.

Box, G. E. P., and J. L. Anderson (1955), Permutation theory in the derivation of robust criteria and the study of departures from assumptions, *Journal of the Royal Statistical Society, B*, **17**, 1–19.

Bradley, J. V. (1968), *Distribution-free Statistical Tests*, Prentice-Hall, Englewood Cliffs, NJ.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Brown, G. W., and A. M. Mood (1948), Homogeneity of several samples, *The American Statistician*, **2**, 22.

Brown, G. W., and A. M. Mood (1951), On median tests for linear hypotheses, 159–166, in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.

Brown, T. S., and J. K. Evans (1986), Hemispheric dominance and recall follwing graphical and tabular presentation of information, *Proceedings of the 1986 Annual Meeting of the Decision Sciences Institute*, **1**, 595–598.

Capon, J. (1961), Asymptotic efficiency of certain locally most powerful rank tests, *Annals of Mathematical Statistics*, **32**, 88–100.

Capon, J. (1965), On the asymptotic efficiency of the Kolmogorov-Smirnov test, *Journal of the American Statistical Association*, **60**, 843–853.

Chacko, V. J. (1963), Testing homogeneity against ordered alternatives, *Annals of Mathematical Statistics*, **34**, 945–956.

Chakraborti, S., and M. M. Desu (1988*a*), A class of distribution-free tests for testing homogeneity against ordered alternatives, *Statistics and Probability Letters*, **6**, 254–256.

Chakraborti, S., and M. M. Desu (1988*b*), Generalization of Mathisen's median test for comparing several treatments with a control, *Communications in Statistics—Simulation and Computation*, **17**, 947–967.

Chakraborti, S., and M. M. Desu (1991), Linear rank tests for comparing treatments with a control when data are subject to unequal patterns of censorship, *Statistica Neerlandica*, **45**, 227–254.

Chakraborti, S., and J. D. Gibbons (1991), One-sided nonparametric comparison of treatments with a standard in the one-way layout, *Journal of Quality Technology*, **23**, 102–106.

Chakraborti, S., and J. D. Gibbons (1992), One-sided comparison of treatments with a standard: Case of the one-way layout with unequal sample sizes, *Journal of Experimental Education*, **60**, 235–242.

Chakraborti, S., and R. Mukerjee (1990), A confidence interval for a measure associated with the comparison of a treatment with a control, *South African Statistical Journal*, **23**, 219–230.

Chakraborti, S., and P. van der Laan (1996), Precedence tests and confidence bounds for complete data, *The Statistician*, **45**, 351–369.

Chakraborti, S., and P. van der Laan (1997), An overview of precedence tests for censored data, *Biometrical Journal*, **39**, 99–116.

Chernoff, H., and E. L. Lehmann (1954), The use of the maximum likelihood estimate in χ^2 tests for goodness of fit, *Annals of Mathematical Statistics*, **25**, 579–589.

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. (1952), The χ^2 test of goodness of fit, *Annals of Mathematical Statistics*, **23**, 315–345.

Cochran, W. G., and G. M. Cox (1957), *Experimental Designs*, John Wiley & Sons, New York.
Cohen, A. J. (1977), *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York.
Conover, W. J. (1967), A k-sample extension of the one-sided two sample Smirnov test statistic, *Annals of Mathematical Statistics*, **38**, 1726–1730.
Conover, W. J. (1972), A Kolmogorov goodness-of-fit test for discontinuous distributions, *Journal of the American Statistical Association*, **67**, 591–596.
Conover, W. J. and R. L. Iman (1981), Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics (C/R: P129–133), *American Statistician*, **35**, 124–129.
Conover, W. J., O. Wehmanen, and F. L. Ramsey (1978), A note on the small-sample power functions for nonparametric tests of location in the double exponential family. *Journal of the American Statistical Association*, **73**, 188–190.
Conover, W. J. (1999), *Practical Nonparametric Statistics*, 3rd ed., John Wiley & Sons, New York.
Cramér, H. (1928), On the composition of elementary errors, *Skandinavisk Aktuarietidskrift*, **11**, 13–74, 141–180.
Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ.
Dallal, G. E., and L. Wilkinson (1986), An analytical approximation to the distribution of Lilliefor's test statistic, *The American Statistician*, **40**, 294–296.
Daniel, W. (1990), *Applied Nonparametric Statistician*, 2nd ed., Houghton Mifflin, Boston.
Darling, D. A. (1957), The Kolmogorov-Smirnov, Cramer-von Mises tests, *Annals of Mathematical Statistics*, **28**, 823–838.
David, F. N. (1947), A power function for tests for randomness in a sequence of alternatives, *Biometrika*, **34**, 335–339.
David, F. N., and D. E. Barton (1958), A test for birth-order effects, *Annals of Human Eugenics*, **22**, 250–257.
David, H. T. (1952), Discrete populations and the Kolmogorov-Smirnov tests, *University of Chicago Statistical Research Center Report* SRC-21103D27.
De Jonge, C., and M. A. J. Van Montfort (1972), The null distribution of Spearman's S when $n = 12$, *Statistica Neerlandica*, **26**, 15–17.
D'Agostino, R. B., and M. A. Stephens, (1986), *Goodness-of-Fit Techniques*, Marcel Dekker, New York.
Dielman, T., C. Lowry and R. Pfaffenberger (1994), A comparison of quantile estimators. *Communications in Statistics–Simulation and Computation*, **23**, 355–371.
Doksum, K. A. (1977), Some graphical methods in statistics: A review and some extensions, *Statistica Neerlandica*, **31**, 53–68.

Donahue, R. M. J. (1999), A note on information seldom reported via the P value, *The American Statistician*, **53**, 303–306.

Drion, E. F. (1952), Some distribution free tests for the difference between two empirical cumulative distributions, *Annals of Mathematical Statistics*, **23**, 563–574.

Dudewicz, E. J., and N. L. Geller (1972), Review of *Nonparametric Statistical Inference* by J. D. Gibbons, *Journal of the American Statistical Association*, **67**, 962–963.

Dunn, O. J. (1964), Multiple comparisons using rank sums, *Technometrics*, **6**, 241–252.

Dunstan, F. D. J., A. B. J. Nix, and J. F. Reynolds (1979), *Statistical Tables*, R.N.D. Publication, Cardiff, United Kingdom.

Duran, B. S. (1976), A survey of nonparametric tests for scale, *Communications in Statistics—Theory and Methods*, **A5**, 1287–1312.

Durbin, J. (1951), Incomplete blocks in ranking experiments, *British Journal of Psychology* (*Statistical Section*), **4**, 85–90.

Durbin, J. (1975), Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentially and tests on spacings, *Biometrika*, **62**, 5–22.

Edgeman, R. L., and R. C. Scott, (1987), Lilliefors's tests for transformed variables, *Revista Brasileira de Probabilidade e Estatistica*, **1**, 101–112.

Edgington, E. S. (1961), Probability table for number of runs of signs or first differences in ordered series, *Journal of the American Statistical Association*, **56**, 156–159.

Elam, R. P. (1988), Morphological changes in adult males from resistance exercise and amino acid supplementation, *Journal of Sports Medicine and Physical Fitness*, **28**, 35–39.

Emerson, J. D., and G. A. Simon (1979), Another look at the sign test when ties are present: The problem of confidence intervals, *The American Statistician*, **33**, 140–142.

Fairley D., and M. Fligner (1987), Linear rank statistics for the ordered alternatives problems, *Communications in Statistics—Theory and Methods*, **16**, 1–16.

Fieller, E. C., and E. S. Pearson (1961), Tests for rank correlation coefficients: II, *Biometrika*, **48**, 29–40.

Fieller, E. C., H. O. Hartley, and E. S. Pearson (1957), Tests for rank correlation coefficients: I, *Biometrika*, **44**, 470–481.

Fisher, N. I. (1983), Graphical methods in nonparametric statistics: A review and annotated bibliography, *International Statistical Review*, **51**, 25–58.

Fisz, M. (1963), *Theory of Probability and Mathematical Statistics*, John Wiley & Sons, New York.

Fix, E., and J. L. Hodges, Jr. (1955), Significance probabilities of the Wilcoxon test, *Annals of Mathematical Statistics*, **26**, 301–312.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Fligner, M. A., and S. W. Rust (1982), A modification of Mood's median test for the generalized Behrens–Fisher problem, *Biometrika*, **69**, 221–226.

Fligner, M. A., and D. A. Wolfe (1976), Some applications of sample analogues to the probability integral transformation and a coverage property, *The American Statistician*, **30**, 78–85.

Fligner, M. A., and D. A. Wolfe (1982), Distribution-free tests for comparing several treatments with a control, *Statistica Neerlandica*, **36**, 119–127.

Franklin, L. A. (1987), A note on approximations and convergence in distribution for Spearman's rank correlation coefficient, *Communications in Statistics—Theory and Methods*, **17**, 55–59.

Franklin, L. A. (1988), The complete exact null distribution of Spearman's rho for $n = 12(1)16$, *Journal of Statistical Computation and Simulation*, **29**, 255–269.

Franklin, L. A. (1989), A note on the Edgeworth approximation to the distribution of Spearman's rho with a correction to Pearson's approximation, *Communications in Statistics—Simulation and Computation*, **18**, 245–252.

Fraser, D. A. S. (1957), *Nonparametric Methods in Statistics*, John Wiley & Sons, New York.

Freund, J. E., and A. R. Ansari (1957), Two-way rank sum tests for variance, *Virginia Polytechnic Institute Technical Report to Office of Ordnance Research and National Science Foundation* 34, Blacksburg, VA.

Friedlin, B., and J. L. Gastwirth, (2000), Should the median test be retired from general use?, *The American Statistician*, **54**, 161–164.

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.

Friedman, M. (1940), A comparison of alternative tests of significance for the problem of m rankings, *Annals of Mathematical Statistics*, **11**, 86–92.

Gart, J. J. (1963), A median test with sequential applications, *Biometrika*, **50**, 55–62.

Gastwirth, J. L. (1965), Percentile modifications of two sample rank tests, *Journal of the American Statistical Association*, **60**, 1127–1141.

Gastwirth, J. L. (1968), The first median test: A two-sided version of the control median test, *Journal of the American Statistical Association*, **63**, 692–706.

Gastwirth, J. L., and J. L. Wang (1988), Control percentile test procedures for censored data, *Journal of Statistical Planning and Inference*, **18**, 267–276.

Giambra, L. M., and R. E. Quilter (1989), Sex differences in sustained attention across the adult life span, *Journal of Applied Psychology*, **74**, 91–95.

Gibbons, J. D. (1964), On the power of two-sample rank tests on the equality of two distribution functions, *Journal of the Royal Statistical Society*, *B*, **26**, 293–304.

Gibbons, J. D. (1973), Comparisons of asymptotic and exact power for percentile modified rank tests, *Sankhyā*, *B*, **35**, 15–24.

Gibbons, J. D. (1982), Fisher's exact test, *Encyclopedia of Statistical Sciences*, **3**, 118–121.
Gibbons, J. D. (1997), *Nonparametric Methods for Quantitative Analysis*, 3rd ed., American Sciences Press, Syracuse, NY.
Gibbons, J. D., and S. Chakraborti (1991), Comparisons of the Mann–Whitney, Student's t, and alternate t tests for means of normal distributions, *Journal of Experimental Education*, **59**, 258–267.
Gibbons, J. D., and J. L. Gastwirth (1966), Small sample properties of percentile modified rank tests, *Johns Hopkins University Department of Statistics Technical Report 60*, Baltimore, MD.
Gibbons, J. D., and J. L. Gastwirth (1970), Properties of the percentile modified rank tests, *Annals of the Institute of Statistical Mathematical Supplement*, **6**, 95–114.
Gibbons, J. D., and J. W. Pratt (1975), *P*-values: Interpretation and methodology, *The American Statistican*, **29**, 20–25.
Gideon, R. A., and D. E. Mueller (1978), Computation of the two-sample Smirnov statistics, *The American Statistician*, **32**, 136–137.
Glasser, G. J., and R. F. Winter (1961), Critical values of the coefficient of rank correlation for testing the hypothesis of independence, *Biometrika*, **48**, 444–448.
Gnedenko, B. V. (1954), Tests of homogeneity of probability distributions in two independent samples (in Russian), *Doklody Akademii Nauk* SSSR, **80**, 525–528.
Goodman, L. A. (1954), Kolmogorov-Smirnov tests for psychological research, *Psychological Bulletin*, **51**, 160–168.
Goodman, L. A., and W. H. Kruskal (1954, 1959, 1963), Measures of association for cross-classifications, *Journal of the American Statistical Association*, **49**, 732–764; **54**, 123–163; **58**, 310–364.
Govindarajulu, Z. (1972), Review of *Nonparametric Statistical Inference* by J. D. Gibbons, *Mathematics Reviews*, **44**, No. 3 (September), No. 3437, 641.
Govindarajulu, Z. (1976), A brief survey of nonparametric statistics, *Communications in Statistics—Theory and Methods*, **A5**, 429–453.
Graubard, B. I., and E. L. Korn (1987), Choices of column scores for testing independence in ordered $2 \times K$ contingency tables, *Biometrics*, **43**, 471–476.
Gumbel, E. J. (1944), Ranges and midranges, *Annals of Mathematical Statistics*, **15**, 414–422.
Gumbel, E. J. (1958), *Statistics of Extremes*, Columbia University Press, New York.
Hackl, P., and W. Katzenbeisser (1984), A note on the power of two-sample tests for dispersion based on exceedance statistics, *Computational Statistics Quarterly*, **1**, 333–341.
Hajék, J. (1969), *Nonparametric Statistics*, Holden-Day, San Francisco.
Hajék, J., and Z. Sidak (1967), *Theory of Rank Tests*, Academic Press, New York.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Harrell, F. E., and C. E. Davis (1982), A new distribution-free quantile estimator, *Biometrika*, **69**, 635–640.

Harter, H. L. (1961), Expected values of normal order statistics, *Biometrika*, **48**, 151–165.

Hartley, H. O. (1942), The probability integral of the range in samples of n observations from the normal population, *Biometrika*, **32**, 301–308.

Hettmansperger, T. P. (1973), A large sample conservative test for location with unknown scale parameters, *Journal of the American Statistical Association*, **68**, 406–408.

Hettmansperger, T. P., and J. S. Malin (1975), A modified Mood test for location with no shape assumptions on the underlying distribution, *Biometrika*, **62**, 527–529.

Hettmansperger, T. P., and R. M. Norton (1987), Tests for patterned alternatives in k-sample problems, *Journal of the American Statistical Association*, **82**, 292–299.

Hilton, J., C. R. Mehta, and N. R. Patel (1994), An algorithm for conducting exact Smirnov tests, *Computational Statistics and Data Analysis*, **17**, 351–361.

Hodges, J. L., Jr. (1958), The significance probability of the Smirnov two-sample test, *Arkiv foer Matematik*, *Astronomi och Fysik*, **3**, 469–486.

Hodges, J. L., Jr., and E. L. Lehmann (1956), The efficiency of some nonparametric competitors of the t-test, *Annals of Mathematical Statistics*, **27**, 324–335.

Hodges, J. L., Jr., and E. L. Lehmann (1970), Deficiency, *Annals of Mathematical Statistics*, **41**, 783–801.

Hoeffding, W. (1951), Optimum nonparametric tests, 83–92, in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California press, Berkeley.

Hoel, P. G. (1962), *Introduction to Mathematical Statistics*, John Wiley & Sons, New York.

Hogg, R. V. (1965), On models and hypotheses with restricted alternatives, *Journal of the American Statistical Association*, **60**, 1153–1162.

Hogg, R. V., and A. T. Craig (1995), *Introduction to Mathematical Statistics*, 5th edition, Prentice Hall, Upper Saddle River, NJ.

Hollander, M., and D. A. Wolfe (1999), Nonparametric Statistical Methods, 2nd ed., John Wiley & Sons, New York.

Hoskin, A. F., D. Yalung-Mathews, and B. A. Carraro (1986), The effect of raising the minimum legal drinking age on fatal crashes in 10 states, *Journal of Safety Research*, **3**, 117–121.

Howard, G. S., C. M. Murphy, and G. E. Thomas (1986), Computer anxiety considerations for design of introductory computer courses, *Proceedings of the 1986 Meetings of the Decision-Sciences Institute*, **1**, 630–632.

Iman, R. L., and W. J. Conover (1981), Rank transformations as a bridge between parametric and nonparametric statistics, *The American Statistician*, **35**, 124–129.

Iman, R. L., and J. M. Davenport (1976), New approximations to the exact distribution of the Kruskal-Wallis test statistic, *Communications in Statistics—Theory and Methods*, **A5**, 1335–1348.

Iman, R. L., D. Quade, and D. A. Alexander (1975), Exact probability levels for the Kruskal-Wallis test, 329–384, in Institute of Mathematical Statistics, ed., *Selected Tables in Mathematical Statistics*, vol. III, American Mathematical Society, Providence, RI.

Iman, R. L. (1982), Graphs for use with the Lilliefors test for normal and exponential distributions, *The American Statistician*, **36**, 109–112.

Johnson, R. A., and K. G. Mehrotra (1971), some c-sample nonparametric tests for ordered alternatives, *Journal of the Indian Statistical Association*, **9**, 8–23.

Johnson, W. D. (1973), Review of *Nonparametric Statistical Inference* by J. D. Gibbons, *Technometrics*, **15**, 421.

Jonckheere, A. R. (1954), A distribution-free k-sample test against ordered alternatives, *Biometrika*, **41**, 133–145.

Jones, M. C. (1993), Review of Nonparametric Statistical Inference, 3rd ed., *Journal of the Royal Statistical Society*, A, **156**, 503.

Kaarsemaker, L., and A. Van Wijngaarden (1952), Tables for use in rank correlation, *Report R 73*, Computation Mathematical Centre, Amsterdam.

Kaarsemaker, L., and A. Van Wijngaarden (1953), Tables for use in rank correlation, *Statistica Neerlandica*, **7**, 41–54.

Kac, M., J. Kiefer, and J. Wolfowitz (1955), On tests of normality and other tests of goodness of fit based on distance methods, *Annals of Mathematical Statistics*, **26**, 189–211.

Kamat, A. R. (1956), A two-sample distribution-free test, *Biometrika*, **43**, 377–387.

Kendall, M. G. (1962), *Rank Correlation Methods*, Hafner Publishing Company, New York.

Kendall, M. G., and J. D. Gibbons (1990), *Rank Correlation Methods*, 5th ed., Edward Arnold, London.

Killeen, T. J., T. P. Hettmansperger, and G. L. Sievers (1972), An elementary theorem on the probability of large deviations, *Annals of Mathematical Statistics*, **43**, 181–192.

Kim, P. J., and R. I. Jennrich (1973), Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov criterion $D_{m,n} (m \leqslant n)$, 79–170, in Institute of Mathematical Statistics, ed., *Selected Tables in Mathematical Statistics*, vol. I, American Mathematical Society, Providence, RI.

Klotz, J. (1962), Nonparametric tests for scale, *Annals of Mathematical Statistics*, **33**, 495–512.

Klotz, J. (1963), Small sample power and efficiency for the one-sample Wilcoxon and normal scores test, *Annals of Mathematical Statistics*, **34**, 624–632.

Klotz, J. (1964), On the normal scores two-sample rank test, *Journal of the American Statistical Association*, **49**, 652–664.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Klotz, J. (1965), Alternative efficiencies for signed rank tests, *Annals of Mathematical Statistics*, **36**, 1759–1766.

Klotz, J. (1972), Review of *Nonparametric Statistical Inference* by J. D. Gibbons, *Biometrics*, **28**, 1148–1149.

Kolmogorov, A. (1933), Sulla determinazione empirica di una legge di distribuzione, *Giornale dell'Instituto Italiano degla Attuari*, **4**, 83–91.

Kolmogorov, A. (1941), Confidence limits for an unknown distribution function, *Annals of Mathematical Statistics*, **12**, 461–463.

Korolyuk, V. S. (1961), On the discrepancy of empiric distributions for the case of two independent samples, *Selected Translations in Mathematical Statistics and Probability*, **1**, 105–121.

Kraft, C. H., and C. van Eeden (1968), *A Nonparametric Introduction to Statistics*, The Macmillan Company, New York.

Kruskal, W. H. (1952), A nonparametric test for the several sample problem, *Annals of Mathematical Statistics*, **23**, 525–540.

Kruskal, W. H. (1958), Ordinal measures of association, *Journal of the American Statistical Association*, **53**, 814–861.

Kruskal, W. H., and W. A. Wallis (1952), Use of ranks in one-criterion analysis of variance, *Journal of the American Statistical Association*, **47**, 583–621; errata, *ibid*., **48**, 907–911.

Laan, P. van der, and J. Prakken (1972), Exact distribution of Durbin's distribution-free test statistic for balanced incomplete block designs, and comparison with the chi-square and F approximation, *Statistica Neerlandica*, **26**, 155–164.

Laan, P. van der, and L. R. Verdooren (1987), Classical analysis of variance methods and nonparametric counterparts, *Biometrical Journal*, **29**, 635–665.

Lachenbruch, P. A. (1992), On the sample size for studies based on McNemar's test, *Statistics in Medicine*, **11**, 1521–1527.

Lachin, J. M. (1992), Power and sample size evaluation for the McNemar test with application to matched case-control studies, *Statistics in Medicine*, **11**, 1239–1251.

Laubscher, N. F., and R. E. Odeh (1976), A confidence interval for the scale parameter based on Sukhatme's two-sample statistic, *Communications in Statistics—Theory and Methods*, **A5**, 1393–1407.

Laubscher, N. F., F. E. Steffens, and E. M. DeLange (1968), Exact critical values for Mood's distribution-free test statistic for dispersion and its normal approximation, *Technometrics*, **10**, 497–508.

Lehmann, E. L. (1953), The power of rank test, *Annals of Mathematical Statistics*, **24**, 23–43.

Lehmann, E. L., and H. J. M. D'Abrera (1998), *Nonparametrics: Statistical Methods Based on Ranks*, Prentice-Hall, Upper Saddle River, NJ.

Levene, H. (1952), On the power function of tests of randomness based on runs up and down, *Annals of Mathematical Statistics*, **23**, 34–56.

Levene, H., and J. Wolfowitz (1944), The covariance matrix of runs up and down, *Annals of Mathematical Statistics*, **15**, 58–69.

Lieberman, G. J., and D. B. Owen (1961), *Tables of the Hypergeometric Probability Distribution*, Stanford University Press, Stanford, CA.

Lienert, G. A. (1978), *Verteilungsfreie Methoden in der Biostatisik*, Band II, Verlag Anton Hain, Meisenheim am Glan, West Germany.

Lilliefors, H. W. (1967), On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association*, **62**, 399–402.

Lilliefors, H. W. (1969), On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown, *Journal of the American Statistical Association*, **64**, 387–389.

Maghsoodloo, S. (1975), Estimates of the quantiles of Kendall's partial rank correlation coefficient and additional quantile estimates, *Journal of Statistical Computation and Simulation*, **4**, 155–164.

Maghsoodloo, S., and L. L. Pallos (1981), Asymptotic behavior of Kendall's partial rank correlation coefficient and additional quantile estimates, *Journal of Statistical Computation and Simulation*, **13**, 41–48.

Mann, H. B., and D. R. Whitney (1947), On a test whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, **18**, 50–60.

Mansfield, E. (1989), Technological creativity: Japan and the United States, *Business Horizons*, **3**(2), 48–53.

Marascuilo, L. A., and M. McSweeney (1977), *Nonparametric and Distribution-Free Methods for the Social Sciences*, Brooks/Cole Publishing Company, Monterey, CA.

Maritz, J. S. (1981), *Distribution-Free Statistical Methods*, Chapman and Hall, London.

Massey, F. J. (1950), A note on the estimation of a distribution function by confidence limits, *Annals of Mathematical Statistics*, **21**, 116–119; correction, *ibid.*, **23**, 637–638 (1952).

Massey, F. J. (1951*a*), A note on a two sample test, *Annals of Mathematical Statistics*, **22**, 304–306.

Massey, F. J. (1951*b*), The distribution of the maximum deviation between two sample cumulative step functions, *Annals of Mathematical Statistics*, **22**, 125–128.

Massey, F. J. (1951*c*), The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American Statistical Association*, **46**, 68–78.

Massey, F. J. (1952), Distribution table for the deviation between two sample cumulatives, *Annals of Mathematical Statistics*, **23**, 435–441.

Mathisen, H. C. (1943), A method of testing the hypotheses that two samples are from the same population, *Annals of Mathematical Statistics*, **14**, 188–194.

McNemar, Q. (1947), Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, **12**, 153–157.

Michaelis, J. (1971), Schwellenwerte des Friedman-tests, *Biometrische Zeitschrift*, **13**, 118–129.
Miller, R. G., Jr. (1966), *Simultaneous Statistical Inference*, McGraw-Hill Book Company, New York.
Miller, R. G., Jr. (1981), *Simultaneous Statistical Inference*, 2nd ed., Springer-Verlag, New York.
Mises, R. von (1931), *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*, F. Deuticke, Leipzig-Wien, 574.
Mood, A. M. (1940), The distribution theory of runs, *Annals of Mathematical Statistics*, **11**, 367–392.
Mood, A. M. (1950), *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, New York, 394–406.
Mood, A. M. (1954), On the asymptotic efficiency of certain nonparametric two-sample tests, *Annals of Mathematical Statistics*, **25**, 514–522.
Mood, A. M., and F. A. Graybill (1963), *Introduction tot the Theory of Statistics*, 2nd ed., McGraw-Hill Book Company, New York.
Mood, A. M., F. A. Graybill, and D. C. Boes (1974), *Introduction to the Theory of Statistics*, 3rd ed., McGraw-Hill Book Company, New York, 504–526.
Moore, B. A. (1986), Review of *Nonparametric Statistical Inference* (2nd ed.) by J. D. Gibbons, *Applied Statistics*, **35**, 215–216.
Moran, P. A. P. (1951), Partial and multiple rank correlation, *Biometrika*, **38**, 26–32.
Moses, L. E. (1963), Rank test for dispersion, *Annals of Mathematical Statistics*, **34**, 973–983.
Mosteller, F. (1941), Note on an application of runs to quality control charts. *Annals of Mathematical Statistics*, **12**, 228–232.
Mosteller, F., and R. E. K. Rourke (1973), *Sturdy Statistics: Nonparametrics and Order Statistics*, Addison-Wesley Publishing Company, Reading, MA.
National Bureau of Standards (1949), *Tables of the Binomial Probability Distribution*, Appl. Math. Ser. 6, U.S. Government Printing Office, Washington, DC.
Neave, H. R. (1981), *Elementary Statistics Tables*, George Allen and Unwin, London.
Nelson, L. S. (1986), Critical values for sums of squared rank differences in Spearman's correlation task, *Journal of Quality Technology*, **18**, 194–196.
Nesbitt, P. D. (1972). Chronic smoking and emotionality, *Journal of Applied Social Psychology*, **2**, 187–196.
Nijsse, M. (1988), Testing the significance of Kendall's τ and Spearman's r_s, *Psychological Bulletin*, **103**, 235–237.
Nikitin, Y. (1995), *Asymptotic efficiency of nonparametric tests*, Cambridge University Press, Cambridge, UK.
Noether, G. E. (1955), On a theorem of Pitman, *Annals of Mathematical Statistics*, **26**, 64–68.

Noether, G. E. (1967), *Elements of Nonparametric Statistics*, John Wiley & Sons, New York.
Noether, G. E. (1972), Review of *Nonparametric Statistical Inference* by J. D. Gibbons, *SIAM Review*, **14**, 346–348.
Noether, G. E. (1973), *Introduction to Statistics: A Nonparametric Approach*, Houghton Mifflin Company, Boston.
Noether, G. E. (1987), Sample size determination for some common nonparametric tests, *Journal of the American Statistical Association*, **82**, 645–647.
Noether, G. E. (1991), *Introduction to Statistics the Nonparametric Way*, Springer-Verlag, New York.
Odeh, R. E. (1971), On Jonckheere's k-sample test against ordered alternatives, *Technometrics*, **13**, 912–918.
Odeh, R. E. (1977), Extended tables of the distribution of Friedman's S-statistic in the two-way layout, *Communications in Statistics—Simulation and Computation*, **B6**, 29–48.
Oden, D. L., R. K. R. Thompson, and D. Premack (1988), Spontaneous matching by infant chimpanzees, *Journal of Experimental Psychology*, **14**, 140–145.
Olejnik, S. F. (1988), Variance heterogeneity: An outcome to explain or a nuisance factor to control, *Journal of Experimental Education*, **56**, 193–197.
Olmstead, P. S. (1946), Distribution of sample arrangements for runs up and down, *Annals of Mathematical Statistics*, **17**, 24–33.
Orban, J., and D. A. Wolfe (1982), A class of distribution-free two-sample tests based on placements, *Journal of the American Statistical Association*, **77**, 666–671.
Otten, A. (1973*a*), Note on the Spearman rank correlation coefficient, *Journal of the American Statistical Association*, **68**, 585.
Otten, A. (1973*b*), The null distribution of Spearman's S when $n = 13(1)16$, *Statistica Neerlandica*, **27**(1), 19–20.
Owen, D. B. (1962), *Handbook of Statistical Tables*, Addision-Wesley Publishing Company, Reading, MA.
Page, E. P. (1963), Ordered hypotheses for multiple treatments: A significance test for linear ranks, *Journal of the American Statistical Association*, **58**, 216–230.
Parzen, E. (1960), *Modern Probability Theory and Its Applications*, John Wiley & 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 reasonably be supposed to have arisen in random sampling, *Philosophical Magazine*, **50**, 157–175.
Pearson, K. (1904), On the theory of contingency and its relation to association and normal correlation, *Draper's Company Research Memoirs Biometric Series I*.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Pettitt, A. N., and M. A. Stephens (1977), The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data, *Technometrics*, **19**, 205–210.
Pirie, W. R. (1979), Computation of the two-sample Smirnov statistics, *The American Statistician*, **33**, 228.
Pothoff, R. F. (1963), Use of the Wilcoxon statistic for a generalized Behrens-Fisher problem, *Annals of Mathematical Statistics*, **34**, 1596–1599.
Pratt, J. W., and J. D. Gibbons (1981), *Concepts of Nonparametric Theory*, Springer-Verlag, New York.
Prvan, T. (1993). Review of Nonparametric Statistical Inference, 3rd ed., *Australian Journal of Statistics*, **35**, 383.
Puri, M. L. (1964), Asymptotic efficiency of a class of c-sample tests, *Annals of Mathematical Statistics*, **35**, 102–121.
Puri, M. L. (1965), Some distribution-free k-sample rank tests for homogeneity against ordered alternatives, *Communications in Pure and Applied Mathematics*, **18**, 51–63.
Puri, M. L., and P. K. Sen (1971), *Nonparametric Methods in Multivariate Analysis*, John Wiley & Sons, New York.
Putter, J. (1955), The treatment of ties in some nonparametric tests, *Annals of Mathematical Statistics*, **26**, 368–386.
Quade, D. (1967), Nonparametric partial correlation, *University of North Carolina, Department of Biostatistics*, *Institute of Statistics Mimeo Series* 526, Chapel Hill, NC.
Quade, D. (1972), Average internal rank correlation, *Technical Report SW 16/72*, Mathematische Centrum, Amsterdam.
Ramsay, P. H. (1989), Critical values for Spearman's rank order correlation, *Journal of Educational Statistics*, **14**, 245–254.
Randles, R. H. (1986), Review of *Nonparametric Statistical Inference* (2nd ed.) by J. D. Gibbons, *Technometrics*, **28**, 275.
Randles, R. H. (2001), On neutral responses (zeros) in the sign test and ties in the Wilcoxon-Mann-Whitney test, *The American Statistician*, **55**, 96–101.
Randles, R. H., and R. V. Hogg (1973), Adaptive distribution-free tests, *Communications in Statistics—Theory and Methods*, **2**, 337–356.
Randles, R. H., and D. A. Wolfe (1979), *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons, New York.
Rijkoort, P. J. (1952), A generalization of Wilcoxon's test, *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, **A55**, 394–404.
Robertson, T., F. T. Wright, and R. L. Dykstra (1988), *Order Restricted Statistical Inference*, John Wiley & Sons, New York.
Rosenbaum, S. (1953), Tables for a nonparametric test of dispersion, *Annals of Mathematical Statistics*, **24**, 663–668.
Rosenbaum, S. (1954), Tables for a nonparametric test of location, *Annals of Mathematical Statistics*, **25**, 146–150.

Rosenbaum, S. (1965), On some two-sample non-parametric tests, *Journal of the American Statistical Association*, **60**, 1118–1126.
Roy, S. N. (1953), On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics*, **24**, 220–238.
Ruben, H. (1954), On the moments of order statistics in samples from normal populations, *Biometrika*, **41**, 200–227.
Runyon, R. P. (1977), *Nonparametric Statistics: A Contemporary Approach*, Addison-Wesley Publishing Company, Reading, MA.
Ryan, T. A., and B. L. Joiner (1976), Normal probability plots and tests for normality, Technical Report, Statistics Department, Pennsylvania State University.
Sackrowitz, H., and E. Samuel-Cahn (1999), P values as random variables, *The American Statistician*, **53**, 326–331.
Sarhan, A. E., and B. G. Greenberg (1962), *Contributions to Order Statistics*, John Wiley & Sons, New York.
Savage, I. R. (1962), *Bibliography of Nonparametric Statistics*, Harvard University Press, Cambridge, MA.
Schlittgen, R. (1979), Use of a median test for a generalized Behrens-Fisher problem, *Metrika*, **26**, 95–103.
Sen, P. K. (1962), On the role of a class of quantile test in some multisample nonparametric problems, *Calcutta Statistical Association Bulletin*, **11**, 125–143.
Sen, P. K. (1964), On some asymptotic properties of a class of nonparametric tests based on the number of rare exceedances, *Annals of the Institute of Statistical Mathematics*, **18**, 319–336.
Shirahata, S. (1980), Rank tests for the k-sample problem with restricted alternatives, *Communications in Statistics—Theory and Methods*, **9**, 1071–1086.
Shorack, G. R. (1967), Testing against order alternatives in model I analysis of variance, normal theory and nonparametric, *Annals of Mathematical Statistics*, **38**, 1740–1752.
Siegel, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, New York.
Siegel, S., and N. J. Castellan (1988), *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed., McGraw-Hill Book Company, New York.
Siegel, S., and J. W. Tukey (1960), A nonparametric sum of ranks procedure for relative spread in unpaired samples, *Journal of the American Statistical Association*, **55**, 429–445; correction, *ibid.,* **56**, 1005 (1961).
Slakter, M. J. (1965), A comparison of the Pearson chi-squared and Kolmogorov goodness-of-fit tests with respect to validity, *Journal of the American Statistical Association*, **60**, 854–858.
Smirnov, N. V. (1935), Uber die Verteilung des allgemeinen Gliedes in der Variationsreihe, *Metron*, **12**, 59–81.
Smirnov, N. V. (1936), Sur la distribution de w^2, *Comptes Rendus*, **202**, 449–452.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Smirnov, N. V. (1939), Estimate of deviation between empirical distribution functions in two independent samples (in Russian), *Bulletin of Moscow University*, **2**, 3–16.
Smirnov, N. V. (1948), Table for estimating the goodness of fit of empirical distributions, *Annals of Mathematical Statistics*, **19**, 279–281.
Smit, C. F. (1980), On a distribution-free Behrens-Fisher test by Hettmansperger & Malin, *Biometrika*, **67**, 241–242.
Somers, R. H. (1959), The rank analogue of product-moment partial correlation and regression, with application to manifold, ordered contingency tables, *Biometrika*, **46**, 241–246.
Sprent, P., and N. C. Smeeton (2000), *Applied Nonparametric Statistical Methods*, 3rd ed., Chapman and Hall, New York.
Stephens, M. A. (1986), Tests based on regression and correlation, Chapter 5, 195–233, in *Goodness-of-fit Techniques* by R. B. D'Agostino and M. A. Stephens (1986), Marcel Dekker, New York.
Streissguth, A. P., D. C. Martin, H. M. Barr, B. M. Sandman, G. L. Kirchner, and B. L. Darby (1984), Intrauterine alcohol and nicotine exposure: Attention and reaction time in 4-year-old children, *Developmental Psychology*, **20**, 533–541.
Struckman-Johnson, C. (1988), Forced sex on dates: It happens to men too, *Journal of Sex Research*, **24**, 234–241.
Stuart, A. (1954), The correlation between variate-values and ranks in samples from a continuous distribution, *British Journal of Statistical Psychology*, **7**, 37–44.
Sukhatme, B. V. (1957), On certain two sample nonparametric tests for variances, *Annals of Mathematical Statistics*, **28**, 188–194.
Sukhatme, S. (1987), Review of *Nonparametric Statistical Inference* (2nd ed.) by J. D. Gibbons, *Journal of the American Statistical Association*, **82**, 953.
Swed, F. S., and C. Eisenhart (1943), Tables for testing randomness of grouping in a sequence of alternatives, *Annals of Mathematical Statistics*, **14**, 66–87.
Tate, M. W., and R. C. Clelland (1957), *Nonparametric and Shortcut Statistics*, Interstate Publishers & Printers, Danville, IL.
Teichroew, D. (1956), Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution, *Annals of Mathematical Statistics*, **27**, 410–426.
Terpstra, T. J. (1952), The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking, *Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen*, **55**, 327–333.
Terry, M. E. (1952), Some rank order tests which are most powerful against specific parametric alternatives, *Annals of Mathematical Statistics*, **23**, 346–366.
Tietjen, G. L., D. K. Kahaner, and R. J. Beckman (1977), Variances and covariances of the normal order statistics for sample sizes 2 to 50, 1–73, in Institute of Mathematical Statistics, ed., *Selected Tables in*

Mathematical Statistics, vol. V, American Mathematical Society, Providence, RI.

Tryon, P. V., and T. P. Hettmansperger (1973), A class of nonparametric tests for homogeneity against ordered alternatives, *Annals of Statistics*, **1**, 1061–1070.

Vollandt, R., and M. Horn (1997), Evaluation of Noether's method of sample size determination for the Wilcoxon-Mann-Whitney test, *Biometrical Journal*, **39**, 823–829.

Waerden, B. L. van der (1952, 1953), Order tests for the two-sample problem and their power, I, II, III, *Indagationes Math. 14 [Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen 55]*, 453–458; *Indag, 15 [Proc. 6]*, 303–310, 311–316; correction, *Indag. 15 [Proc. 56]*, 80.

Waerden, B. L. van der, and E. Nievergelt (1956), *Tafeln zum Vergleich Zweier Stichprobenmittels X-test und Zeichentest*, Springer-Verlag, OHG, Berlin.

Wald, A., and J. Wolfowitz (1940), On a test whether two samples are from the same population, *Annals of Mathematical Statistics*, **11**, 147–162.

Wald, A., and J. Wolfowitz (1943), An exact test for randomness in the non-parametric case based on serial correlation, *Annals of Mathematical Statistics*, **14**, 378–388.

Walsh, J. E. (1949*a*), Applications of some significance tests for the median which are valid under very general conditions, *Journal of the American Statistical Association*, **44**, 342–355.

Walsh, J. E. (1949*b*), Some significance tests for the median which are valid under very general conditions, *Annals of Mathematical Statistics*, **20**, 64–81.

Walsh, J. E. (1962), *Handbook of Nonparametric Statistics, I: Investigation of Randomness, Moments, Percentiles, and Distributions*, Van Nostrand Company, New York.

Walsh, J. E. (1965), *Handbook of Nonparametric Statistics, II: Results for Two and Several Sample Problems, Symmetry and Extremes*, Van Nostrand Company, New York.

Walsh, J. E. (1968), *Handbook of Nonparametric Statistics, III: Analysis of Variance*, Van Nostrand Company, New York.

Webber, R. A. (1990), Ethics gap, *Wharton Alumni Magazine*, Summer, 39–40.

Westenberg, J. (1948), Significance test for median and interquartile range in samples from continuous populations of any form, *Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen*, **51**, 252–261.

Wilcoxon, F. (1945), Individual comparisons by ranking methods, *Biometrics*, **1**, 80–83.

Wilcoxon, F. (1947), Probability tables for individual comparisons by ranking methods, *Biometrics*, **3**, 119–122.

Wilcoxon, F. (1949), *Some Rapid Approximate Statistical Procedures*, American Cyanamid Company, Stanford Research Laboratories, Stanford, CA.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Wilcoxon, F., S. K. Katti, and R. A. Wilcox (1972), Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test, 171–259, in Institute of Mathematical Statistics, ed., *Selected Tables in Mathematical Statistics*, vol. I, American Mathematical Society, Providence, RI.

Wilk, M. B., and R. Gnanadesikan. (1968), Probability plotting methods for the analysis of data, *Biometrika*, **55**, 1–17.

Wilks, S. S. (1948), Order statistics, *Bulletin of the American Mathematical Society*, **54**, 6–50.

Wilks, S. S. (1962), *Mathematical Statistics*, John Wiley & Sons, New York.

Wolfowitz, J. (1944*a*), Asymptotic distribution of runs up and down, *Annals of Mathematical Statistics*, **15**, 163–172.

Wolfowitz, J. (1944*b*), Note on runs of consecutive elements, *Annals of Mathematical Statistics*, **15**, 97–98.

Wolfowitz, J. (1949), Nonparametric statistical inference, 93–113, in J. Neyman, ed., *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.

Young, D. H. (1973), A note on some asymptotic properties of the precedence test and applications to a selection problem concerning quantiles, *Sankhyā*, B, **35**, 35–44.

Zar, J. H. (1972), Significance testing of the Spearman rank correlation coefficient, *Journal of the American Statistical Association*, **67**, 578–580.

Ziegel, E. (1988), Review of *Nonparametric Statistical Inference* (2nd ed.) by J. D. Gibbons, *Technometrics*, **30**, 457.

Ziegel, E. (1993), Review of *Nonparametric Statistical Inference*, 3rd ed., *Technometrics*, **35**, 239–240.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Index

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016