Kenneth J. Berry
Janis E. Johnston
Paul W. Mielke Jr.

A Chronicle of Permutation Statistical Methods

1920–2000, and Beyond

EXTRA
MATERIALS
extras.springer.com

Springer

A Chronicle of Permutation Statistical Methods

Kenneth J. Berry • Janis E. Johnston •
Paul W. Mielke Jr.

A Chronicle of Permutation Statistical Methods

1920–2000, and Beyond

Springer

Kenneth J. Berry
Department of Sociology
Colorado State University
Fort Collins, CO
USA

Janis E. Johnston
U.S. Government
Alexandria, VA
USA

Paul W. Mielke Jr.
Department of Statistics
Colorado State University
Fort Collins, CO
USA

Additional material to this book can be downloaded from http://extra.springer.com.

ISBN 978-3-319-02743-2 ISBN 978-3-319-02744-9 (eBook)
DOI 10.1007/978-3-319-02744-9
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014935885

© Springer International Publishing Switzerland 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*For our families: Nancy T. Berry,
Ellen E. Berry, Laura B. Berry,
Lindsay A. Johnston, James B. Johnston,
Roberta R. Mielke, William W. Mielke,
Emily (Mielke) Spear, and Lynn (Mielke)
Basila.*

Preface

The stimulus for this volume on the historical development of permutation statistical methods from 1920 to 2000 was a 2006 Ph.D. dissertation by the second author on ranching in Colorado in which permutation methods were extensively employed [695]. This was followed by an invited overview paper on permutation statistical methods in *Wiley Interdisciplinary Reviews: Computational Statistics*, by all three authors in 2011 [117]. Although a number of research monographs and textbooks have been published on permutation statistical methods, few have included much historical material, with the notable exception of Edgington and Onghena in the fourth edition of their book on *Randomization Tests* published in 2007 [396]. In addition, David provided a brief history of the beginnings of permutation statistical methods in a 2008 publication [326], which was preceded by a more technical and detailed description of the structure of permutation tests by Bell and Sen in 1984 [93]. However, none of these sources provides an extensive historical account of the development of permutation statistical methods.

As Stephen Stigler noted in the opening paragraph of his 1999 book on *Statistics on the Table: The History of Statistical Concepts and Methods*:

> [s]tatistical concepts are ubiquitous in every province of human thought. they are more likely to be noticed in the sciences, but they also underlie crucial arguments in history, literature, and religion. As a consequence, the history of statistics is broad in scope and rich in diversity, occasionally technical and complicated in structure, and never covered completely [1321, p. 1].

This book emphasizes the historical and social context of permutation statistical methods, as well as the motivation for the development of selected permutation tests. The field is broadly interpreted and it is notable that many of the early pioneers were major contributors to, and may be best remembered for, work in other disciplines and areas. Many of the early contributors to the development of permutation methods were trained for other professions such as mathematics, economics, agriculture, the military, or chemistry. In more recent times, researchers from atmospheric science, biology, botany, computer science, ecology, epidemiology, environmental health, geology, medicine, psychology, and sociology have made significant contributions to the advancement of permutation statistical methods. Their common characteristic was an interest in, and capacity to use, quantitative methods on problems judged to be important in their respective disciplines.

The purpose of this book is to chronicle the birth and development of permutation statistical methods over the approximately 80-year period from 1920 to 2000. As to what the state of permutation methods will be 80 years in the future—one can only guess. Not even our adult children will live to see the permutation methods of that day. As for ourselves, we have to deal with the present and the past. It is our hope in this writing that knowledge of the past will help the reader to think critically about the present. Those who write intellectual history, as Hayden White maintained, "do not build up knowledge that others might use, they generate a discourse about the past" (White, quoted in Cohen [267, pp. 184–185]). Although the authors are not historians, they are still appreciative of the responsibility historians necessarily assume when trying to accurately, impartially, and objectively interpret the past. Moreover, the authors are acutely aware of the *1984* Orwellian warning that "Who controls the past . . . controls the future" [1073, p. 19]. The authors are also fully cognizant that there are the records of the past, then there is the interpretation of those records. The gap between them is a source of concern. As Appleby, Hunt, and Jacob noted in *Telling the Truth About History*, "[a]t best, the past only dimly corresponds to what the historians say about it" [28, p. 248]. In writing this book, the authors were reminded of the memorable quote by Walter Sellar and Robert Yeatman, the authors of *1066 and All That: A Memorable History of England*: "History is not what you thought. *It is what you can remember*" [1245, p. vii].[1] In researching the development of permutation methods, the authors constantly discovered historical events of which they were not aware, remembered events they thought they had forgotten, and often found what they thought they remembered was incorrect. Debates as to how to present historical information about the development of permutation methods will likely be prompted by this volume. What is not up for debate is the impact that permutation methods have had on contemporary statistical methods. Finally, as researchers who have worked in the field of statistics for many years, the authors fondly recall a sentient quote by Karl Pearson:

> I do feel how wrongful it was to work for so many years at statistics and neglect its history [1098, p. 1].

A number of books and articles detailing the history of statistics have been written, but there is little coverage of the historical development of permutation methods. While many of the books and articles have briefly touched on the development of permutation methods, none has been devoted entirely to the topic. Among the many important sources on the history of probability and statistics, a few have served the authors well, being informative, interesting, or both. Among these we count *Natural Selection, Heredity and Eugenics: Selected Correspondence of R.A. Fisher with Leonard Darwin and Others* and *Statistical Inference and Analysis: Selected Correspondence of R.A. Fisher* by J.H. Bennett [96, 97]; "A history of statistics in the social sciences" by V. Coven [289]; *A History of Inverse Probability from Thomas Bayes to Karl Pearson* by A.I. Dale [310]; *Games, Gods,*

[1] Emphasis in the original.

and Gambling: The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era by F.N. David [320]; “Behavioral statistics: An historical perspective” by A.L. Dudycha and L.W. Dudycha [361]; “A brief history of statistics in three and one-half chapters” by S.E. Fienberg [428]; *The Making of Statisticians* edited by J. Gani [493]; *The Empire of Chance: How Probability Changed Science and Everyday Life* by G. Gigerenzer, Z. Swijtink, T.M. Porter, and L. Daston [512]; *The Emergence of Probability* and *The Taming of Chance* by I. Hacking [567, 568]; *History of Probability and Statistics and Their Applications Before 1750* and *A History of Mathematical Statistics from 1750 to 1930* by A. Hald [571,572]; “The method of least squares and some alternatives: Part I,” “The method of least squares and some alternatives: Part II,” “The method of least squares and some alternatives: Part III,” “The method of least squares and some alternatives: Part IV,” “The method of least squares and some alternatives: Addendum to Part IV,” “The method of least squares and some alternatives: Part V,” and “The method of least squares and some alternatives: Part VI” by H.L. Harter [589–595]; *Statisticians of the Centuries* edited by C.C. Heyde and E. Seneta [613]; *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present* edited by N.L. Johnson and S. Kotz [691]; *Bibliography of Statistical Literature: 1950–1958*, *Bibliography of Statistical Literature: 1940–1949*, and *Bibliography of Statistical Literature: Pre 1940* by M.G. Kendall and A.G. Doig [743–745].

Also, *Studies in the History of Statistics and Probability* edited by M.G. Kendall and R.L. Plackett [747]; *Creative Minds, Charmed Lives: Interviews at Institute for Mathematical Sciences, National University of Singapore* edited by L.Y. Kiang [752]; “A bibliography of contingency table literature: 1900 to 1974” by R.A. Killion and D.A. Zahn [754]; *The Probabilistic Revolution* edited by L. Krüger, L. Daston, and M. Heidelberger [775]; *Reminiscences of a Statistician: The Company I Kept* and *Fisher, Neyman, and the Creation of Classical Statistics* by E.L. Lehmann [814, 816]; *Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge* by D. MacKenzie [863]; *The History of Statistics in the 17th and 18th Centuries Against the Changing Background of Intellectual, Scientific and Religious Thought* edited by E.S. Pearson [1098]; *Studies in the History of Statistics and Probability* edited by E.S. Pearson and M.G. Kendall [1103]; *The Rise of Statistical Thinking, 1820–1900* by T.M. Porter [1141]; *Milestones in Computer Science and Information Technology* by E.D. Reilly [1162]; *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* by D. Salsburg [1218]; *Bibliography of Nonparametric Statistics* by I.R. Savage [1225]; *Theory of Probability: A Historical Essay* by O.B. Sheynin [1263]; *American Contributions to Mathematical Statistics in the Nineteenth Century, Volumes 1 and 2*, *The History of Statistics: The Measurement of Uncertainty Before 1900*, and *Statistics on the Table: The History of Statistical Concepts and Methods* by S.M. Stigler [1318–1321], *Studies in the History of Statistical Method* by H.M. Walker [1409], and the 44 articles published by various authors under the title “Studies in the history of probability and statistics” that appeared in *Biometrika* between 1955 and 2000.

In addition, the authors have consulted myriad addresses, anthologies, articles, autobiographies, bibliographies, biographies, books, celebrations, chronicles, collections, commentaries, comments, compendiums, compilations, conversations, correspondences, dialogues, discussions, dissertations, documents, essays, eulogies, encyclopedias, festschrifts, histories, letters, manuscripts, memoirs, memorials, obituaries, remembrances, reports, reviews, speeches, summaries, synopses, theses, tributes, web sites, and various other sources on the contributions of individual statisticians to permutation methods, many of which are listed in the references at the end of the book.

No preface to a chronicle of the development of permutation statistical methods would be complete without acknowledging the major contributors to the field, some of whom contributed theory, others methods and algorithms, and still others promoted permutation methods to new audiences. At the risk of slighting someone of importance, in the early years from 1920 to 1939 important contributions were made by Thomas Eden, Ronald Fisher, Roy Geary, Harold Hotelling, Joseph Irwin, Jerzy Neyman, Edwin Olds, Margaret Pabst, Edwin Pitman, Bernard Welch, and Frank Yates. Later, the prominent names were Bernard Babington Smith, George Box, Meyer Dwass, Eugene Edgington, Churchill Eisenhart, Alvan Feinstein, Leon Festinger, David Finney, Gerald Freeman, Milton Friedman, Arthur Ghent, John Haldane, John Halton, Wassily Hoeffding, Lawrence Hubert, Maurice Kendall, Oscar Kempthorne, William Kruskal, Erich Lehmann, Patrick Leslie, Henry Mann, M. Donal McCarthy, Cyrus Mehta, Nitin Patel, Henry Scheffé, Cedric Smith, Charles Spearman, Charles Stein, John Tukey, Abraham Wald, Dirk van der Reyden, W. Allen Wallis, John Whitfield, Donald Whitney, Frank Wilcoxon, Samuel Wilks, and Jacob Wolfowitz. More recently, one should recognize Alan Agresti, Brian Cade, Herbert David, Hugh Dudley, David Freedman, Phillip Good, Peter Kennedy, David Lane, John Ludbrook, Bryan Manly, Patrick Onghena, Fortunato Pesarin, Jon Richards, and Cajo ter Braak.

Acknowledgments. The authors wish to thank the editors and staff at Springer-Verlag. A very special thanks to Federica Corradi Dell'Acqua, Assistant Editor, Statistics and Natural Language Processing, who guided the project through from beginning to end; this book would not have been written without her guidance and oversight. We also wish to thank Norm Walsh who answered all our LaTeX questions. We are grateful to Roberta Mielke who read the entire manuscript and made many helpful comments, and Cristi MacWaters, Interlibrary Loan Coordinator at Morgan Library, Colorado State University, who retrieved many of the manuscripts we needed. Finally, we wish to thank Steve and Linda Jones, proprietors of the Rainbow Restaurant, 212 West Laurel Street, Fort Collins, Colorado, for their gracious hospitality; the bulk of this book was written at table 20 in their restaurant adjacent to the campus of Colorado State University.

Fort Collins, CO — Kenneth J. Berry
Alexandria, VA — Janis E. Johnston
Fort Collins, CO — Paul W. Mielke Jr.
August 2013

Acronyms

2-D	Two-dimensional
3-D	Three-dimensional
AAAS	American Association for the Advancement of Science
ACM	Association for Computing Machinery
AEC	Atomic Energy Commission
ALGOL	Algorithmic computer language
AMAP	Approximate Multivariate Association Procedure
ANOVA	Analysis of variance
APL	A programming language
ARE	Asymptotic relative efficiency
ARPAnet	Advanced Research Projects Agency network
ASCC	Automatic sequence controlled calculator
ASR	Automatic send and receive
BAAS	British Association for the Advancement of Science
BASIC	Beginners All-Purpose Symbolic Instruction Code
BBS	Bernard Babington Smith
BIT	BInary digiT
CCNY	City College of New York
CBS	Columbia Broadcasting System
CDC	Control Data Corporation
CDF	Cumulative distribution function
CEEB	College Entrance Examination Board
CF	Correction factor (analysis of variance)
CIT	California Institute of Technology
CM	Correction factor
COBOL	Common business oriented language
CPU	Central processing unit
CSM	Company sergeant major
CTR	Computing Tabulating Recording Corporation
DARPA	Defense Advanced Research Projects Agency
DEC	Digital Equipment Corporation
DHSS	Department of Health and Social Security
DOD	Department of Defense
DOE	The design of experiments (Fisher)

ECDF	Empirical cumulative distribution function
ECST	Exact chi-squared test
EDA	Exploratory data analysis
EDSAC	Electronic delay storage automatic calculator
EEG	Electroencephalogram
EM	Engineer of mines
EMAP	Exact multivariate association procedure
ENIAC	Electronic numerical integrator and computer
EPA	Environmental Protection Agency
ETH	Eidgenössische Technische Hochschule
ETS	Educational Testing Service
FEPT	Fisher exact probability test
FFT	Fast Fourier transform
FLOPS	Floating operations per second
FNS	Food and Nutrition Service
FORTRAN	Formula Translation
FRS	Fellow of the Royal Society
GCHQ	Government Communications Head Quarters
Ge	Germanium
GE	General electric
GL	Generalized logistic (distribution)
GOF	Goodness of fit
GPD	Generalized Pareto distribution
GUI	Graphical user interface
IAS	Institute for Advanced Study (Princeton)
IBM	International Business Machines (Corporation)
ICI	Imperial Chemical Industries
IEEE	Institute of Electrical and Electronics Engineering
IML	Integer Matrix Library
IMS	Institute of Mathematical Statistics
IP	Internet protocol
IRBA	Imagery Randomized Block Analysis
KΣ	Kappa sigma (fraternity)
LAD	Least absolute deviation (regression)
LANL	Los Alamos National Laboratory
LASL	Los Alamos Scientific Laboratory
LEO	Lyons Electronic Office
LGP	Librascope General Purpose
LINC	Laboratory Instrument Computer
LLNL	Lawrence Livermore National Laboratory
LSED	Least sum of Euclidean distances
MANIAC	Mathematical analyzer, numerical integrator and computer
MANOVA	Multivariate analysis of variance
MCM	Micro computer machines
MIT	Massachusetts Institute of Technology

MITS	Micro Instrumentation Telemetry Systems
MPP	Massively parallel processing
MRBP	Multivariate randomized block permutation procedures
MRPP	Multi-response permutation procedures
MS	Mean square (analysis of variance)
MSPA	Multivariate sequential permutation analyses
MT	Mersenne Twister
MXH	Multivariate extended hypergeometric
NBA	National Basketball Association
NBS	National Bureau of Standards
NCAR	National Center for Atmospheric Research
NCR	National Cash Register Company
NFL	National Football League
NHSRC	National Homeland Security Research Center
NIST	National Institute of Standards and Technology
NIT	National Institutes of Health
NRC	National Research Council
NSF	National Science Foundation
NSFNET	National Science Foundation NETwork
NYU	New York University
OBE	Order of the British Empire
OECD	Organization for Economic Cooperation and Development
OLS	Ordinary least squares (regression)
ONR	Office of Naval Research
ORACLE	Oak Ridge Automatic Computer and Logical Engine
OSRD	Office of Scientific Research and Development
PC	Personal computer
PDP	Programmed data processor
PET	Personal Electronic Transactor (Commodore PET)
ΦΚΘ	Phi kappa theta (fraternity)
PISA	Programme for International Student Assessment
PKU	Phenylketonuria
PRNG	Pseudo random number generator
PSI	Statisticians in the Pharmaceutical Industry
RAF	Royal Air Force
RAND	Research and Development (Corporation)
RE	Random error
RIDIT	Relative to an identified distribution
SAGE	Semi-Automatic Ground Environment
SAT	Scholastic aptitude test
SFMT	SIMD-Oriented Fast Mersenne Twister
SIAM	Society for Industrial and Applied Mathematics
SIMD	Single instruction [stream], multiple data [stream]
SiO_2	Silicon oxide
SK	Symmetric kappa (distribution)

SLC	Super Little Chip
SNL	Sandia National Laboratories
SPSS	Statistical Package for the Social Sciences
SREB	Southern Regional Education Board
SRG	Statistical Research Group (Columbia University)
SRI	Stanford Research Institute
SS	Sum of squares (analysis of variance)
SSN	Spanish Supercomputing Network
SUN	Stanford University Network
TAOCP	The Art of Computer Programming
TRS	Tandy Radio Shack
TCP	Transmission Control Protocol
UCLA	University of California, Los Angeles
UNIVAC	Universal Automatic Computer
USDA	United States Department of Agriculture
WMW	Wilcoxon–Mann–Whitney two-sample rank-sum test

Contents

Introduction 1

Permutation statistical methods are a paradox of old and new. While permutation methods pre-date many traditional parametric statistical methods, only recently have permutation methods become part of the mainstream discussion regarding statistical testing. Permutation statistical methods follow a permutation model whereby a test statistic is computed on the observed data, then (1) the observed data are permuted over all possible arrangements of the observations—an exact permutation test, (2) the observed data are used for calculating the exact moments of the underlying discrete permutation distribution and the moments are fitted to an associated continuous distribution—a moment-approximation permutation test, or (3) the observed data are permuted over a random subset of all possible arrangements of the observations—a resampling-approximation permutation test [977, pp. 216–218].

1.1 Overview of This Chapter

This first chapter begins with a brief description of the advantages of permutation methods from statisticians who were, or are, advocates of permutation tests, followed by a description of the methods of permutation tests including exact, moment-approximation, and resampling-approximation permutation tests. The chapter continues with an example that contrasts the well-known Student t test and results from exact, moment-approximation, and resampling-approximation permutation tests using historical data. The chapter concludes with brief overviews of the remaining chapters.

Permutation tests are often described as the gold standard against which conventional parametric tests are tested and evaluated. Bakeman, Robinson, and Quera remarked that "like Read and Cressie (1988), we think permutation tests represent the standard against which asymptotic tests must be judged" [50, p. 6]. Edgington and Onghena opined that "randomization tests . . . have come to be recognized by many in the field of medicine as the 'gold standard' of statistical tests for randomized experiments" [396, p. 9]; Friedman, in comparing tests of significance

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9_1,
© Springer International Publishing Switzerland 2014

for m rankings, referred to an exact permutation test as "the correct one" [486, p. 88]; Feinstein remarked that conventional statistical tests "yield reasonably reliable approximations of the more exact results provided by permutation procedures" [421, p. 912]; and Good noted that Fisher himself regarded randomization as a technique for validating tests of significance, i.e., making sure that conventional probability values were accurate [521, p. 263].

Early statisticians understood well the value of permutation statistical tests even during the period in which the computationally-intensive nature of the tests made them impractical. Notably, in 1955 Kempthorne wrote that "[t]ests of significance in the randomized experiment have frequently been presented by way of normal law theory, whereas their validity stems from randomization theory" [719, p. 947] and

> [w]hen one considers the whole problem of experimental inference, that is of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using method of inference other randomization analysis [719, p. 966].

In 1966 Kempthorne re-emphasized that "the proper way to make tests of significance in the simple randomized experiments is by way of the randomization (or permutation) test" [720, p. 20] and "in the randomized experiment one should, logically, make tests of significance by way of the randomization test" [720, p. 21].[1] Similarly, in 1959 Scheffé stated that the conventional analysis of variance F test "can often be regarded as a good approximation to a permutation [randomization] test, which is an exact test under a less restrictive model" [1232, p. 313]. In 1968 Bradley indicated that "eminent statisticians have stated that the randomization test is the truly correct one and that the corresponding parametric test is valid only to the extent that it results in the same statistical decision" [201, p. 85].

With the advent of high-speed computing, permutation tests became more practical and researchers increasingly appreciated the benefits of the randomization model. In 1998, Ludbrook and Dudley stated that "it is our thesis that the randomization rather than the population model applies, and that the statistical procedures best adapted to this model are those based on permutation" [856, p. 127], concluding that "statistical inferences from the experiments are valid only under the randomization model of inference" [856, p. 131].

In 2000, Bergmann, Ludbrook, and Dudley, in a cogent analysis of the Wilcoxon–Mann–Whitney two-sample rank-sum test, observed that "the only accurate form of the Wilcoxon–Mann–Whitney procedure is one in which the exact permutation null distribution is compiled for the actual data" [100, p. 72] and concluded:

> [o]n theoretical grounds, it is clear that the only infallible way of executing the [Wilcoxon–Mann–Whitney] test is to compile the null distribution of the rank-sum statistic by exact permutation. This was, in effect, Wilcoxon's (1945) thesis and it provided the theoretical basis for his [two-sample rank-sum] test [100, p. 76].

[1] The terms "permutation test" and "randomization test" are often used interchangeably.

1.2 Two Models of Statistical Inference

Essentially, two models of statistical inference coexist: the population model and the permutation model; see for further discussion, articles by Curran-Everett [307], Hubbard [663], Kempthorne [721], Kennedy [748], Lachin [787], Ludbrook [849, 850], and Ludbrook and Dudley [854]. The population model, formally proposed by Jerzy Neyman and Egon Pearson in 1928 [1035, 1036], assumes random sampling from one or more specified populations. Under the population model, the level of statistical significance that results from applying a statistical test to the results of an experiment or a survey corresponds to the frequency with which the null hypothesis would be rejected in repeated random samplings from the same specified population(s). Because repeated sampling of the true population(s) is usually impractical, it is assumed that the sampling distribution of the test statistics generated under repeated random sampling conforms to an assumed, conjectured, hypothetical distribution, such as the normal distribution.

The size of a statistical test, e.g., 0.05, is the probability under a specified null hypothesis that repeated outcomes based on random samples of the same size are equal to or more extreme than the observed outcome. In the population model, assignment of treatments to subjects is viewed as fixed with the stochastic element taking the form of an error that would vary if the experiment was repeated [748]. Probability values are then calculated based on the potential outcomes of conceptual repeated draws of these errors. The model is sometimes referred to as the "conditional-on-assignment" model, as the distribution used for structuring the test is conditional on the treatment assignment of the observed sample; see for example, a comprehensive and informative 1995 article by Peter Kennedy in *Journal of Business & Economic Statistics* [748].

The permutation model was introduced by R.A. Fisher in 1925 [448] and further developed by R.C. Geary in 1927 [500], T. Eden and F. Yates in 1933 [379], and E.J.G. Pitman in 1937 and 1938 [1129–1131]. Permutation tests do not refer to any particular statistical tests, but to a general method of determining probability values. In a permutation statistical test the only assumption made is that experimental variability has caused the observed result. That assumption, or null hypothesis, is then tested. The smaller the probability, the stronger is the evidence against the assumption [648]. Under the permutation model, a permutation test statistic is computed for the observed data, then the observations are permuted over all possible arrangements of the observations and the test statistic is computed for each equally-likely arrangement of the observed data [307]. For clarification, an ordered sequence of n exchangeable objects $(\omega_1, \ldots, \omega_n)$ yields $n!$ equally-likely arrangements of the n objects, *vide infra*. The proportion of cases with test statistic values equal to or more extreme than the observed case yields the probability of the observed test statistic. In contrast to the population model, the assignment of errors to subjects is viewed as fixed, with the stochastic element taking the form of the assignment of treatments to subjects for each arrangement [748]. Probability values are then calculated according to all outcomes associated with assignments of treatments to subjects for each case. This model is sometimes referred to as the "conditional-on-errors" model, as the distribution used for structuring the test is conditional on the individual errors drawn for the observed sample; see for example, a 1995 article by Peter Kennedy [748].

Exchangeability

A sufficient condition for a permutation test is the exchangeability of the random variables. Sequences that are independent and identically distributed (i.i.d.) are always exchangeable, but so is sampling without replacement from a finite population. However, while i.i.d. implies exchangeability, exchangeability does not imply i.i.d. [528, 601, 758]. Diaconis and Freedman present a readable discussion of exchangeability using urns and colored balls [346].

More formally, variables $X_1, X_2, \ldots, X_n$ are exchangeable if

$$P\left[\bigcap_{i=1}^{n}(X_i \leq x_i)\right] = P\left[\bigcap_{i=1}^{n}(X_i \leq x_{c_i})\right] ,$$

where $x_1, x_2, \ldots, x_n$ are n observed values and $\{c_1, c_2, \ldots, c_n\}$ is any one of the $n!$ equally-likely permutations of $\{1, 2, \ldots, n\}$ [1215].

1.3 Permutation Tests

Three types of permutation tests are common: exact, moment-approximation, and resampling-approximation permutation tests. While the three types are methodologically quite different, all three approaches are based on the same specified null hypothesis.

1.3.1 Exact Permutation Tests

Exact permutation tests enumerate all equally-likely arrangements of the observed data. For each arrangement, the desired test statistic is calculated. The obtained data yield the observed value of the test statistic. The probability of obtaining the observed value of the test statistic, or a more extreme value, is the proportion of the enumerated test statistics with values equal to or more extreme than the value of the observed test statistic. As sample sizes increase, the number of possible arrangements can become very large and exact methods become impractical. For example, permuting two small samples of sizes $n_1 = n_2 = 20$ yields

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(20 + 20)!}{(20!)^2} = 137{,}846{,}528{,}820$$

different arrangements of the observed data.

1.3.2 Moment-Approximation Permutation Tests

The moment-approximation of a test statistic requires computation of the exact moments of the test statistic, assuming equally-likely arrangements of the observed data. The moments are then used to fit a specified distribution. For example, the first three exact moments may be used to fit a Pearson type III distribution. Then, the Pearson type III distribution approximates the underlying discrete permutation distribution and provides an approximate probability value. For many years moment-approximation permutation tests provided an important intermediary approximation when computers lacked both the speed and the storage for calculating exact permutation tests. More recently, resampling-approximation permutation tests have largely replaced moment-approximation permutation tests, except when either the size of the data set is very large or the probability of the observed test statistic is very small.

1.3.3 Resampling-Approximation Permutation Tests

Resampling-approximation permutation tests generate and examine a Monte Carlo random subset of all possible equally-likely arrangements of the observed data. In the case of a resampling-approximation permutation test, the probability of obtaining the observed value of the test statistic, or a more extreme value, is the proportion of the resampled test statistics with values equal to or more extreme than the value of the observed test statistic [368, 649]. Thus, resampling permutation probability values are computationally quite similar to exact permutation tests, but the number of resamplings to be considered is decided upon by the researcher rather than by considering all possible arrangements of the observed data. With sufficient resamplings, a researcher can compute a probability value to any accuracy desired. Read and Cressie [1157], Bakeman, Robinson, and Quera [50], and Edgington and Onghena [396, p. 9] described permutation methods as the "gold standard" against which asymptotic methods must be judged. Tukey took it one step further, labeling resampling permutation methods the "platinum standard" of permutation methods [216, 1381, 1382].[2]

1.3.4 Compared with Parametric Tests

Permutation tests differ from traditional parametric tests based on an assumed population model in several ways.

[2] In a reversal Tukey could not have predicted, at the time of this writing gold was trading at $1,775 per troy ounce, while platinum was only $1,712 per troy ounce [275].

1. Permutation tests are data dependent, in that all the information required for analysis is contained within the observed data set; see a 2007 discussion by Mielke and Berry [965, p. 3].[3]
2. Permutation tests do not assume an underlying theoretical distribution; see a 1983 article by Gabriel and Hall [489].
3. Permutation tests do not depend on the assumptions associated with traditional parametric tests, such as normality and homogeneity; see articles by Kennedy in 1995 [748] and Berry, Mielke, and Mielke in 2002 [162].[4]
4. Permutation tests provide probability values based on the discrete permutation distribution of equally-likely test statistic values, rather than an approximate probability value based on a conjectured theoretical distribution, such as a normal, chi-squared, or F distribution; see a 2001 article by Berry, Johnston, and Mielke [117].
5. Whereas permutation tests are suitable when a random sample is obtained from a designated population, permutation tests are also appropriate for nonrandom samples, such as are common in biomedical research; see discussions by Kempthorne in 1977 [721], Gabriel and Hall in 1983 [489], Bear in 1995 [88], Frick in 1998 [482], Ludbrook and Dudley in 1998 [856], and Edgington and Onghena in 2007 [396, pp. 6–8].
6. Permutation tests are appropriate when analyzing entire populations, as permutation tests are not predicated on repeated random sampling from a specified population; see discussions by Ludbrook and Dudley in 1998 [856], Holford in 2003 [638], and Edgington and Onghena in 2007 [396, pp. 1–8].
7. Permutation tests can be defined for any selected test statistic; thus, researchers have the option of using a wide variety of test statistics, including the majority of statistics commonly utilized in traditional statistical approaches; see discussions by Mielke and Berry in 2007 [965].
8. Permutation tests are ideal for very small data sets, when conjectured, hypothetical distribution functions may provide very poor fits; see a 1998 article by Ludbrook and Dudley [856].
9. Appropriate permutation tests are resistant to extreme values, such as are common in demographic data, e.g., income, age at first marriage, number of children, and so on; see a discussion by Mielke and Berry in 2007 [965, pp. 52–53] and an article by Mielke, Berry, and Johnston in 2011 [978]. Consequently, the need for any data transformation is mitigated in the permutation context and in general is not recommended, e.g., square root, logarithmic, the use of

[3] Echoing Fisher's argument that inference must be based solely on the data at hand [460], Haber refers to data dependency as "the data at hand principle" [565, p. 148].

[4] Barton and David noted that it is desirable to make the minimum of assumptions, since, witness the oft-cited Bertrand paradox [163], that the assumptions made will often prejudice the conclusions reached [83, p. 455].

rank-order statistics,[5] and the choice of a distance function, in particular, may be very misleading [978].

10. Permutation tests provide data-dependent statistical inferences only to the actual experiment or survey that has been performed, and are not dependent on a contrived super population; see for example, discussions by Feinstein in 1973 [421] and Edgington and Onghena in 2007 [396, pp. 7–8].

1.3.5 The Bootstrap and the Jackknife

This chronicle is confined to permutation methods, although many researchers consider that permutation methods, bootstrapping, and the jackknife are closely related. Traditionally, jackknife (leave-one-out) methods have been used to reduce bias in small samples, calculate confidence intervals around parameter estimates, and test hypotheses [789, 876, 1376], while bootstrap methods have been used to estimate standard errors in cases where the distribution of the data is unknown [789]. In general, permutation methods are considered to be more powerful than either the bootstrap or (possibly) the jackknife approaches [789].

While permutation methods and bootstrapping both involve computing simulations, and the rejection of the null hypothesis occurs when a common test statistic is extreme under both bootstrapping and permutation, they are conceptually and mechanically quite different. On the other hand, they do have some similarities, including equivalence in an asymptotic sense [358, 1189]. The two approaches differ in their distinct sampling methods. In resampling, a "new" sample is obtained by drawing the data without replacement, whereas in bootstrapping a "new" sample is obtained by drawing from the data with replacement [748, 1189]. Thus, bootstrapping and resampling are associated with sampling with and without replacement, respectively. Philip Good has been reported as saying that the difference between permutation tests and bootstrap tests is that "[p]ermutations test hypotheses concerning distributions; bootstraps test hypotheses concerning parameters."

Specifically, resampling is a data-dependent procedure, dealing with all finite arrangements of the observed data, and based on sampling without replacement. In contrast, bootstrapping involves repeated sampling from a finite population that conceptually yields an induced infinite population based on sampling with replacement. In addition, when bootstrapping is used with small samples it is necessary to make complex adjustments to control the risk of error; see for example, discussions by Hall and Wilson in 1991 [577], Efron and Tibshirani in 1993 [402], and Westfall and Young, also in 1993 [1437]. Finally, the bootstrap distribution may be viewed as an unconditional approximation to the null distribution of the

[5] Rank-order statistics were among the earliest permutation tests, transforming the observed data into ranks, e.g., from smallest to largest. While they were an important step in the history of permutation tests, modern computing has superseded the need for rank-order tests in the majority of cases.

test statistic, while the resampling distribution may be viewed as a conditional distribution of the test statistic [1189].

In 1991 Donegani argued that it is preferable to compute a permutation test based on sampling without replacement (i.e., resampling) than with replacement (i.e., bootstrap), although, as he noted, the two techniques are asymptotically equivalent [358]. In a thorough comparison and analysis of the two methods, he demonstrated that (1) the bootstrap procedure is "bad" for small sample sizes or whenever the alternative is close to the null hypothesis and (2) resampling tests should be used in order to take advantage of their flexibility in the choice of a distance criteria [358, p. 183].

In 1988 Tukey stated that the relationship between permutation procedures, on the one hand, and bootstrap and jackknife procedures, on the other hand, is "far from close" [1382]. Specifically, Tukey listed four major differences between bootstrap and jackknife procedures, which he called "resampling," and permutation methods, which he called "rerandomization" [1382].

1. Bootstrap and jackknife procedures need not begin until the data is collected. Rerandomization requires planning before the data collection is specified.
2. Bootstrap and jackknife procedures play games of omission of units with data already collected. Rerandomization plays games of exchange of treatments, while using all numerical results each time.
3. Bootstrap and jackknife procedures apply to experiences as well as experiments. Rerandomization only applies to randomized experiments.
4. Bootstrap and jackknife procedures give one only a better approximation to a desired confidence interval. Rerandomization gives one a "platinum standard" significance test, which can be extended in simple cases—by the usual devices—to a "platinum standard" confidence interval.

Thus, bootstrapping remains firmly in the conditional-on-assignment tradition, assuming that the true error distribution can be approximated by a discrete distribution with equal probability attached to each of the cases [850]. On the other hand, permutation tests view the errors as fixed in repeated samples [748]. Finally, some researchers have tacitly conceived of permutation methods in a Bayesian context. Specifically, this interpretation amounts to a primitive Bayesian analysis where the prior distribution is the assumption of equally-likely arrangements associated with the observed data, and the posterior distribution is the resulting data-dependent distribution of the test statistic induced by the prior distribution.

1.4 Student's *t* Test

Student's pooled t test [1331] for two independent samples is a convenient vehicle to illustrate permutation tests and to compare a permutation test with its parametric counterpart. As a historical note, Student's 1908 publication used z for the test statistic, and not t. The first mention of t appeared in a letter from William Sealy Gosset ("Student") to R.A. Fisher in November of 1922. It appears that the decision to change from z to t originated with Fisher, but the choice of the letter t was due to Student. Eisenhart [408] and Box [196] provide historical commentaries on the transition from Student's *z* test to Student's *t* test.

Student's pooled *t* test for two independent samples is well-known, familiar to most researchers, widely used in quantitative analyses, and elegantly simple. The pooled *t* test evaluates the mean difference between two independent random samples. Under the null hypothesis, H_0: $\mu_1 = \mu_2$, Student's pooled *t* test statistic is defined as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \ ,$$

where the standard error of the sampling distribution of differences between two independent sample means is given by

$$s_{\bar{x}_1 - \bar{x}_2} = \left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{n_1 + n_2}{n_1 n_2} \right) \right]^{1/2} ,$$

μ_1 and μ_2 denote the hypothesized population means, $\bar{x}_1$ and $\bar{x}_2$ denote the sample means, s_1^2 and s_2^2 denote the sample variances, and *t* follows Student's *t* distribution with $n_1 + n_2 - 2$ degrees of freedom, assuming the data samples are from independent normal distributions with equal variances.

1.4.1 An Exact Permutation *t* Test

Exact permutation tests are based on all possible arrangements of the observed data. For the two-sample *t* test, the number of permutations of the observed data is given by

$$M = \frac{N!}{n_1!\, n_2!} \ ,$$

where $N = n_1 + n_2$.

Let x_{ij} denote the *i*th observed score in the *j*th independent sample, $j = 1, 2$ and $i = 1, \ldots, n_j$, let t_o denote the Student *t* statistic computed on the observed data, and let t_k denote the Student *t* statistic computed on each permutation of the observed data for $k = 1, \ldots, M$. For the first permutation of the observed data set, interchange x_{13} and x_{12}, compute t_1, and compare t_1 with t_o. For the second permutation, interchange x_{12} and x_{22}, compute t_2, and compare t_2 with t_o. Continue the process for $k = 1, \ldots, M$.

To illustrate the exact permutation procedure, consider two independent samples of $n_1 = n_2 = 3$ observations and let $\{x_{11}, x_{21}, x_{31}\}$ denote the $n_1 = 3$ observations in Sample 1 and $\{x_{12}, x_{22}, x_{32}\}$ denote the $n_2 = 3$ observations in Sample 2. Table 1.1 depicts the

Table 1.1 Illustrative $M = 20$ permutations of $N = 6$ observations in two independent samples with $n_1 = n_2 = 3$

	Sample 1			Sample 2			
Permutation	1	2	3	1	2	3	t
1	x_{11}	x_{21}	x_{31}	x_{12}	x_{22}	x_{32}	t_1
2	x_{11}	x_{21}	x_{12}	x_{31}	x_{22}	x_{32}	t_2
3	x_{11}	x_{21}	x_{22}	x_{31}	x_{12}	x_{32}	t_3
4	x_{11}	x_{21}	x_{32}	x_{31}	x_{12}	x_{22}	t_4
5	x_{11}	x_{31}	x_{12}	x_{21}	x_{22}	x_{32}	t_5
6	x_{11}	x_{31}	x_{22}	x_{21}	x_{12}	x_{32}	t_6
7	x_{11}	x_{31}	x_{32}	x_{21}	x_{12}	x_{22}	t_7
8	x_{11}	x_{12}	x_{22}	x_{21}	x_{31}	x_{32}	t_8
9	x_{11}	x_{12}	x_{32}	x_{21}	x_{31}	x_{22}	t_9
10	x_{11}	x_{22}	x_{32}	x_{21}	x_{31}	x_{12}	t_{10}
11	x_{21}	x_{31}	x_{12}	x_{11}	x_{22}	x_{32}	t_{11}
12	x_{21}	x_{31}	x_{22}	x_{11}	x_{12}	x_{32}	t_{12}
13	x_{21}	x_{31}	x_{32}	x_{11}	x_{12}	x_{22}	t_{13}
14	x_{21}	x_{12}	x_{22}	x_{11}	x_{31}	x_{32}	t_{14}
15	x_{21}	x_{12}	x_{32}	x_{11}	x_{31}	x_{22}	t_{15}
16	x_{21}	x_{22}	x_{32}	x_{11}	x_{31}	x_{12}	t_{16}
17	x_{31}	x_{12}	x_{22}	x_{11}	x_{21}	x_{32}	t_{17}
18	x_{31}	x_{12}	x_{32}	x_{11}	x_{21}	x_{22}	t_{18}
19	x_{31}	x_{22}	x_{32}	x_{11}	x_{21}	x_{12}	t_{19}
20	x_{12}	x_{22}	x_{32}	x_{11}	x_{21}	x_{31}	t_{20}

$$M = \frac{6!}{3!\,3!} = 20$$

arrangements of $n_1 = n_2 = 3$ observations in each of the two independent samples where $t_{\text{o}} = t_1$, the subscripts denote the original position of each observation in either Sample 1 or Sample 2, and the position of the observation in Table 1.1 on either the left side of the table in Sample 1 or the right side of the table in Sample 2 indicates the placement of the observation after permutation. The exact two-sided probability (P) value is then given by

$$P = \frac{\text{number of } |t_k| \text{ values} \geq |t_{\text{o}}|}{M} \qquad \text{for } k = 1, \ldots, M \ .$$

1.4.2 A Moment-Approximation *t* Test

Moment-approximation permutation tests filled an important gap in the development of permutation statistical methods. Prior to the advent of modern computers, exact tests were impossible to compute except for extremely small samples, and even resampling-approximation permutation tests were limited in the number of random permutations of the data possible, thus yielding too few places of accuracy for research purposes.

A moment-approximation permutation test is based, for example, on the first three exact moments of the underlying discrete permutation distribution, yielding the exact mean, variance, and skewness, i.e., μ_x, σ_x^2, and γ_x. Computational details for the exact moments are given in Sect. 4.15 of Chap. 4. An approximate probability value is obtained by fitting the exact moments to the associated Pearson type III distribution, which is completely characterized by the first three moments, and integrating the obtained Pearson type III distribution.

1.4.3 A Resampling-Approximation *t* Test

When M is very large, exact permutation tests are impractical, even with high-speed computers, and resampling-approximation permutation tests become an important alternative. Resampling-approximation tests provide more precise probability values than moment-approximation tests and are similar in structure to exact tests, except that only a random sample of size L selected from all possible permutations, M, is generated, where L is usually a large number to guarantee accuracy to a specified number of places. For instance, $L = 1{,}000{,}000$ will likely ensure three places of accuracy [696]. The resampling two-sided approximate probability value is then given by

$$\hat{P} = \frac{\text{number of } |t_k| \text{ values} \geq |t_{\text{o}}|}{L} \qquad \text{for } k = 1, \ldots, L \; .$$

1.5 An Example Data Analysis

The English poor laws, the relief expenditure act, and a comparison of two English counties provide vehicles to illustrate exact, moment-approximation, and resampling-approximation permutation tests.

The English Poor Laws

Up until the Reformation, it was considered a Christian duty in England to undertake the seven corporal works of mercy. In accordance with Matthew 25:32–46, Christians were to feed the hungry, give drink to the thirsty, welcome a stranger, clothe the naked, visit the sick, visit the prisoner, and bury the dead. After the Reformation and the establishment of the Church of England, many of these precepts were neglected, the poor were left without adequate assistance, and it became necessary to regulate relief of the poor

(continued)

by statute. The Poor Laws passed during the reign of Elizabeth I played a determining role in England's system of welfare, signaling a progression from private charity to a welfare state, where care of the poor was embodied in law. Boyer [198] provides an exhaustive description of the historical development of the English Poor Laws.

In 1552, Parish registers of the poor were introduced to ensure a well-documented official record, and in 1563, Justices of the Peace were empowered to raise funds to support the poor. In 1572, it was made compulsory that all people pay a poor tax, with those funds used to help the deserving poor. In 1597, Parliament passed a law that each parish appoint an Overseer of the Poor who calculated how much money was needed for the parish, set the poor tax accordingly, collected the poor rate from property owners, dispensed either food or money to the poor, and supervised the parish poor house. In 1601, the Poor Law Act was passed by Parliament, which brought together all prior measures into one legal document. The act of 1601 endured until the Poor Law Amendment Act was passed in 1834.

Consider an example data analysis utilizing Student's pooled two-sample t test based on historical parish-relief expenditure data from the 1800s [697]. To investigate factors that contributed to the level of relief expenditures, Boyer [198] assembled a data set comprised of a sample of 311 parishes in 20 counties in the south of England in 1831. The relief expenditure data were obtained from Blaug [172].[6] Table 1.2 contains the 1831 per capita relief expenditures, in shillings, for 36 parishes in two counties: Oxford and Hertford. For this example, the data were rounded to four places.

The relief expenditure data from Oxford and Hertford counties are listed in Table 1.2. Oxford County consisted of 24 parishes with a sample mean relief of $\bar{x}_1 = 20.28$ shillings and a sample variance of $s_1^2 = 58.37$ shillings. Hertford County consisted of 12 parishes with a sample mean relief of $\bar{x}_2 = 13.47$ shillings and a sample variance of $s_2^2 = 37.58$ shillings. A conventional two-sample t test yields $t_o = +2.68$ and, with $24 + 12 - 2 = 34$ degrees of freedom, a two-sided approximate probability value of $\hat{P} = .0113$. Although there are

$$M = \frac{36!}{24!\,12!} = 1{,}251{,}677{,}700$$

possible arrangements of the observed data and an exact permutation test is therefore not practical, it is not impossible. For the Oxford and Hertford relief expenditure

[6] The complete data set is available in several formats at the Cambridge University Press site: http://uk.cambridge.org/resources/0521806631.

Table 1.2 Average per capita relief expenditures for Oxford and Hertford counties in shillings: 1831

Oxford County				Hertford County	
Parish	Expenditure	Parish	Expenditure	Parish	Expenditure
1	20.3619	13	25.4683	1	27.9748
2	29.0861	14	12.5632	2	6.4173
3	14.9318	15	13.2780	3	10.4841
4	24.1232	16	27.3030	4	10.0057
5	18.2075	17	29.6055	5	9.7699
6	20.7287	18	13.6132	6	15.8665
7	8.1195	19	11.3714	7	19.3424
8	14.0201	20	21.5248	8	17.1452
9	18.4248	21	20.9408	9	13.1342
10	34.5466	22	11.5952	10	10.0420
11	16.0927	23	18.2355	11	15.0838
12	24.6166	24	37.8809	12	6.3985

data in Table 1.2, an exact permutation analysis yields a two-sided probability value of $P = 10{,}635{,}310/1{,}251{,}677{,}700 = 0.0085$.

A moment-approximation permutation analysis of the Oxford and Hertford relief expenditure data in Table 1.2 based on the Pearson type III distribution, yields a two-sided approximate probability value of $\hat{P} = 0.0100$.

Finally, a resampling analysis of the Oxford and Hertford relief expenditure data based on $L = 1{,}000{,}000$ random arrangements of the observed data in Table 1.2, yields 8,478 calculated t values equal to or more extreme than the observed value of $t_o = +2.68$, and a two-sided approximate probability value of $\hat{P} = 8{,}478/1{,}000{,}000 = 0.0085$.

1.6 Overviews of Chaps. 2–6

Chapters 2–6 describe the birth and development of statistical permutation methods. Chapter 2 covers the period from 1920 to 1939; Chap. 3, the period from 1940 to 1959; Chap. 4, the period from 1960 to 1979; and Chap. 5, the period from 1980 to 2000. Chapter 6 looks beyond the year 2000, summarizing the development of permutation methods from 2001 to 2010. Following Chap. 6 is a brief epilogue summarizing the attributes that distinguish permutation statistical methods from conventional statistical methods.

Chapter 2: 1920–1939

Chapter 2 chronicles the period from 1920 to 1939 when the earliest discussions of permutation methods appeared in the literature. In this period J. Spława-Neyman, R.A. Fisher, R.C. Geary, T. Eden, F. Yates, and E.J.G. Pitman laid the foundations of permutation methods as we know them today. As is evident in this period, permutation methods had their roots in agriculture and, from the beginning, were widely recognized as the gold standard against which conventional methods could be verified and confirmed.

In 1923 Spława-Neyman introduced a permutation model for the analysis of field experiments [1312], and in 1925 Fisher calculated an exact probability using the binomial distribution [448]. Two years later in 1927, Geary used an exact analysis to support the use of asymptotic methods for correlation and regression [500], and in 1933 Eden and Yates used a resampling-approximation permutation approach to validate the assumption of normality in an agricultural experiment [379].

In 1935, Fisher's well-known hypothesized experiment involving "the lady tasting tea" was published in the first edition of *The Design of Experiments* [451]. In 1936, Fisher used a shuffling technique to demonstrate how a permutation test works [453], and in the same year Hotelling and Pabst utilized permutation methods to calculate exact probability values for the analysis of rank data [653].

In 1937 and 1938, Pitman published three seminal articles on permutation methods. The first article dealt with permutation methods in general, with an emphasis on the two-sample test; the second article with permutation methods as applied to bivariate correlation; and the third article with permutation methods as applied to a randomized blocks analysis of variance [1129–1131].

In addition to laying the foundations for permutation tests, the 1920s and 1930s were also periods in which tools to ease the computation of permutation tests were developed. Probability tables provided exact values for small samples, rank tests simplified the calculations, and desktop calculators became more available. Importantly, statistical laboratories began to appear in the United States in the 1920s and 1930s, notably at the University of Michigan and Iowa State College of Agriculture (now, Iowa State University). These statistical centers not only resulted in setting the foundations for the development of the computing power that would eventually make permutation tests feasible, they also initiated the formal study of statistics as a stand-alone discipline.

Chapter 3: 1940–1959

Chapter 3 explores the period between 1940 and 1959 with attention to the continuing development of permutation methods. This period may be considered as a bridge between the early years where permutation methods were first conceptualized and the next period, 1960–1979, in which gains in computer technology provided the necessary tools to successfully employ specific permutation tests.

Between 1940 and 1959, the work on establishing permutation statistical methods that began in the 1920s continued. In the 1940s, researchers applied known permutation techniques to create tables of exact probability values for small samples, among them tables for 2 × 2 contingency tables; the Spearman and Kendall rank-order correlation coefficients; the Wilcoxon, Mann–Whitney, and Festinger two-sample rank-sum tests; and the Mann test for trend.

Theoretical work, driven primarily by the computational challenges of calculating exact permutation probability values, was also completed during this period. Instead of the focus being on new permutation tests, however, attention turned to developing more simple alternatives to do calculations by converting data to rank-order statistics. Examples of rank tests that were developed between 1940 and 1959 include non-parametric randomization tests, exact tests for randomness based on serial correlation, and tests of significance when the underlying probability distribution is unknown.

While this theoretical undertaking continued, other researchers worked on developing practical non-parametric rank tests. Key among these tests were the Kendall rank-order correlation coefficient, the Kruskal–Wallis one-way analysis of variance rank test, the Wilcoxon and Mann–Whitney two-sample rank-sum tests, and the Mood median test.

Chapter 4: 1960–1979

Chapter 4 surveys the development of permutation methods in the period between 1960 and 1979 that was witness to dramatic improvements in computer technology, a process that was integral to the further development of permutation statistical methods. Prior to 1960, computers were based on vacuum tubes[7] and were large, slow, expensive, and availability was severely limited. Between 1960 and 1979 computers increasingly became based on transistors and were smaller, faster, more affordable, and more readily available to researchers. As computers became more accessible to researchers, work on permutation tests continued with much of the focus of that work driven by computer limitations in speed and storage.

During this period, work on permutation methods fell primarily into three categories: writing algorithms that efficiently generated permutation sequences; designing exact permutation analogs for existing parametric statistics; and, for the first time, developing statistics specifically designed for permutation methods. Numerous algorithms were published in the 1960s and 1970s with a focus on increasing the speed and efficiency of the routines for generating permutation sequences. Other researchers focused on existing statistics, creating permutation counterparts for well-known conventional statistics, notably the Fisher exact probability test for 2×2 contingency tables, the Pitman test for two independent samples, the F test for randomized block designs, and the chi-squared test for goodness of fit. The first procedures designed specifically for permutation methods, multi-response permutation procedures (MRPP), appeared during this period.

[7] The diode and triode vacuum tubes were invented in 1906 and 1908, respectively, by Lee de Forest.

Chapter 5: 1980–2000

Chapter 5 details the development of permutation methods during the period 1980 to 2000. It is in this period that permutation tests may be said to have arrived. One measure of this arrival was the expansion in the coverage of permutation tests, branching out from the traditional coverage areas in computer technology and statistical journals, and into such diverse subject areas as anthropology, atmospheric science, biomedical science, psychology, and environmental health. A second measure of the arrival of permutation statistical methods was the sheer number of algorithms that continued to be developed in this period, including the development of a pivotal network algorithm by Mehta and Patel in 1980 [919]. Finally, additional procedures designed specifically for permutation methods, multivariate randomized block permutation (MRBP) procedures, were published in 1982 by Mielke and Iyer [984].

This period was also home to the first books that dealt specifically with permutation tests, including volumes by Edgington in 1980, 1987 and 1995 [392–394], Hubert in 1987 [666], Noreen in 1989 [1041], Good in 1994 and 1999 [522–524], Manly in 1991 and 1997 [875, 876], and Simon in 1997 [1277], among others. Permutation versions of known statistics continued to be developed in the 1980s and 1990s, and work also continued on developing permutation statistical tests that did not possess existing parametric analogs.

Chapter 6: Beyond 2000

Chapter 6 describes permutation methods after the year 2000, an era in which permutation tests have become much more commonplace. Computer memory and speed issues that hampered early permutation tests are no longer factors and computers are readily available to virtually all researchers. Software packages for permutation tests now exist for well-known statistical programs such as StatXact, SPSS, Stata, and SAS. A number of books on permutation methods have been published in this period, including works by Chihara and Hesterberg in 2011, Edgington and Onghena in 2007 [396], Good in 2000 and 2001 [525–527], Lunneborg in 2000 [858], Manly in 2007 [877], Mielke and Berry in 2001 and 2007 [961, 965], and Pesarin and Salmaso in 2010 [1122].

Among the many permutation methods considered in this period are analysis of variance, linear regression and correlation, analysis of clinical trials, measures of agreement and concordance, rank tests, ridit analysis, power, and Bayesian hierarchical analysis. In addition, permutation methods expanded into new fields of inquiry, including animal research, bioinformatics, chemistry, clinical trials, operations research, and veterinary medicine.

The growth in the field of permutations is made palpable by a search of The Web of Science® using the key word "permutation." Between 1915 and 1959, the key word search reveals 43 journal articles. That number increases to 540 articles for the period between 1960 and 1979 and jumps to 3,792 articles for the period between 1980 and 1999. From 2000 to 2010, the keyword search for permutation results in 9,259 journal articles.

Epilogue

A brief coda concludes the book. Chapter 2 contains a description of the celebrated "lady tasting tea" experiment introduced by Fisher in 1935 [451, pp. 11–29], which is the iconic permutation test. The Epilogue returns full circle to the lady tasting tea experiment, analyzing the original experiment to summarize the attributes that distinguish permutation tests from conventional tests in general.

Researchers early on understood the superiority of permutation tests for calculating exact probability values. These same researchers also well understood the limitations of trying to calculate exact probability values. While some researchers turned to developing asymptotic solutions for calculating probability values, other researchers remained focused on the continued development of permutation tests. This book chronicles the search for better methods for calculating permutation tests, the development of permutation counterparts for existing parametric statistical tests, and the development of separate, unique permutation tests.

1920–1939 2

The second chapter of *A Chronicle of Permutation Statistical Methods* is devoted to describing the earliest permutation tests and the statisticians that developed them. Examples of these early tests are provided and, in many cases, include the original data. The chapter begins with a brief overview of the development of permutation methods in the 1920s and 1930s and is followed by an in-depth treatment of selected contributions. The chapter concludes with a brief discussion of the early threads in the permutation literature that proved to be important as the field progressed and developed from the early 1920s to the present.

2.1 Overview of This Chapter

The 1920s and 1930s ushered in the field of permutation statistical methods. Several important themes emerged in these early years. First was the use of permutation methods to evaluate statistics based on normal theory. Second was the considerable frustration expressed with the difficulty of the computations on which exact permutation methods were based. Third was the widespread reluctance to substitute permutation methods for normal-theory methods, regarding permutation tests as a valuable device, but not as replacements for existing statistical tests. Fourth was the use of moments to approximate the discrete permutation distribution, as exact computations were too cumbersome except for the very smallest of samples. Fifth was the recognition that a permutation distribution could be based on only the variable portion of the sample statistic, thereby greatly reducing the number of calculations required. Sixth was an early reliance on recursion methods to generate successive values of the test statistic. And seventh was a fixation on the use of levels of significance, such as $\alpha = 0.05$, even when the exact probability value was available from the discrete permutation distribution.

The initial contributions to permutation methods were made by J. Spława-Neyman, R.A. Fisher, and R.C. Geary in the 1920s [448, 500, 1312]. Neyman's 1923 article foreshadowed the use of permutation methods, which were developed

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9_2,

© Springer International Publishing Switzerland 2014

by Fisher while at the Rothamsted Experimental Station. In 1927, Geary was the first to use an exact permutation analysis to evaluate and demonstrate the utility of asymptotic approaches. In the early 1930s T. Eden and F. Yates utilized permutation methods to evaluate conventional parametric methods in an agricultural experiment, using a random sample of all permutations of the observed data comprised of measurements on heights of Yeoman II wheat shoots [379]. This was perhaps the first example of the use of resampling techniques in an experiment. The middle 1930s witnessed three articles emphasizing permutation methods to generate exact probability values for 2 × 2 contingency tables by R.A. Fisher, F. Yates, and J.O. Irwin [452,674,1472]. In 1926 Fisher published an article on "The arrangement of field experiments" [449] in which the term "randomization" was apparently used for the first time [176, 323]. In 1935 Fisher compared the means of randomized pairs of observations by permutation methods using data from Charles Darwin on *Zea mays* plantings [451], and in 1936 Fisher described a card-shuffling procedure for analyzing data that offered an alternative approach to permutation statistical tests [453].

In 1936 H. Hotelling and M.R. Pabst utilized permutation methods to circumvent the assumption of normality and for calculating exact probability values for small samples of rank data [653], and in 1937 M. Friedman built on the work of Hotelling and Pabst to investigate the use of rank data in the ordinary analysis of variance [485]. In 1937 B.L. Welch compared the normal theory of Fisher's variance-ratio z test (later, Snedecor's F test) with permutation-version analyses of randomized block and Latin square designs [1428], and in 1938 Welch used an exact permutation test to address tests of homogeneity for the correlation ratio, η^2 [1429]. Egon Pearson was highly critical of permutation methods, especially the permutation methods of Fisher, and in 1937 Pearson published an important critique of permutation methods with special attention to the works of Fisher on the analysis of Darwin's *Zea mays* data and Fisher's thinly-veiled criticism of the coefficient of racial likeness developed by Pearson's famous father, Karl Pearson [1093].

In 1937 and 1938 E.J.G. Pitman published three seminal articles on permutation tests in which he examined permutation versions of two-sample tests, bivariate correlation, and randomized blocks analysis of variance [1129–1131]. Building on the work of Hotelling and Pabst in 1936, E.G. Olds used permutation methods to generate exact probability values for Spearman's rank-order correlation coefficient in 1938 [1054], and in that same year M.G. Kendall incorporated permutation methods in the construction of a new measure of rank-order correlation based on the difference between the sums of concordant and discordant pairs [728]. Finally, in 1939 M.D. McCarthy argued for the use of permutation methods as first approximations before considering the data by means of an asymptotic distribution.

2.2 Neyman–Fisher–Geary and the Beginning

Although precursors to permutation methods based on discrete probability values were common prior to 1920 [396, pp. 13–15], it was not until the early 1920s that statistical tests were developed in forms that are recognized today as permutation methods. The 1920s and 1930s were critical to the development of permutation methods because it was during this nascent period that permutation methods were first conceptualized and began to develop into a legitimate statistical approach. The beginnings are founded in three farsighted publications in the 1920s by J. Spława-Neyman, R.A. Fisher, and R.C. Geary.[1]

2.2.1 Spława-Neyman and Agricultural Experiments

In 1923 Jerzy Spława-Neyman introduced a permutation model for the analysis of agricultural field experiments. This early paper used permutation methods to compare and evaluate differences among several crop varieties [1312].

J. Spława-Neyman

Jerzy Spława-Neyman earned an undergraduate degree from the University of Kharkov (later, Maxim Gorki University[2]) in mathematics in 1917 and the following year was a docent at the Institute of Technology, Kharkov. He took his first job as the only statistician at the National Institute of Agriculture in Bydgoszcz in northern Poland and went on to receive a Ph.D. in mathematics from the University of Warsaw in 1924 with a dissertation, written in Bydgoszcz, on applying the theory of probability to agricultural experiments [817, p. 161]. It was during this period that he dropped the "Spława" from his surname, resulting in the more commonly-recognized Jerzy Neyman. Constance Reid, Spława-Neyman's biographer, explained that Neyman published his early papers under the name Spława-Neyman, and that the word Spława refers to Neyman's family coat of arms and was a sign of nobility [1160, p. 45]. Spława-Neyman is used here because the 1923 paper was published under that name.

After a year of lecturing on statistics at the Central College of Agriculture in Warsaw and the Universities of Warsaw and Krakow, Neyman was sent by the Polish government to University College, London, to study statistics with Karl Pearson [817, p. 161]. Thus it was in 1925 that Neyman moved to England and, coincidentally, began a decade-long association with Egon Pearson, the son of Karl Pearson. That collaboration eventually yielded

(continued)

[1] For an enlightened discussion of the differences and similarities between Neyman and Fisher and their collective impact on the field of statistics, see a 1966 article by Stephen Fienberg and Judith Tanur in *International Statistical Review* [430] and also E.L. Lehmann's remarkable last book, published posthumously in 2011, on *Fisher, Neyman, and the Creation of Classical Statistics* [816].

[2] Maxim Gorki (Maksim Gorky) is a pseudonym for Aleksei Maksimovich Peshkov (1868–1936), Russian short-story writer, novelist, and political activist.

the formal theory of tests of hypotheses and led to Neyman's subsequent invention of confidence intervals [431].

Neyman returned to his native Poland in 1927, remaining there until 1934 whereupon he returned to England to join Egon Pearson at University College, London, as a Senior Lecturer and then Reader. In 1938 Neyman received a letter from Griffith C. Evans, Chair of the Mathematics Department at the University of California at Berkeley, offering Neyman a position teaching probability and statistics in his department. Neyman accepted the offer, moved to Berkeley, and in 1955 founded the Department of Statistics. Neyman formally retired from Berkeley at the age of 66 but at the urging of his colleagues, was permitted to serve as the director of the Statistical Laboratory as Emeritus Professor, remaining an active member of the Berkeley academic community for 40 years. In 1979 Neyman was elected Fellow of the Royal Society.[3] As Lehmann and Reid related, Neyman spent the last days of his life in the hospital with a sign on the door to his room that read, "Family members only," and the hospital staff were amazed at the size of Jerzy's family [817, p. 192]. Jerzy Spława-Neyman F.R.S. passed away in Oakland, California, on 5 August 1981 at the age of 87 [252, 431, 581, 727, 814, 816, 817, 1241].

A brief story will illustrate a little of Neyman's personality and his relationship with his graduate students, of which he had many during his many years at the University of California at Berkeley.

A Jerzy Neyman Story

In 1939, Jerzy Neyman was teaching in the mathematics department at the University of California, Berkeley. Famously, one of the first year doctoral students, George B. Dantzig, arrived late to class, and observing two equations on the chalk-board, assumed they were homework problems and wrote them down. He turned in his homework a few days later apologizing for the delay, noting that these problems had been more difficult than usual. Six weeks later, Dantzig and his wife were awakened early on a Sunday morning by a knock

(continued)

[3] The Royal Society is a fellowship of the world's most eminent scientists and is the oldest scientific society in continuous existence. The society was founded on 28 November 1660 when a group of 12 scholars met at Gresham College and decided to found "a Colledge for the Promoting of Physico-Mathematicall Experimentall Learning" and received a Royal Charter on 5 December 1660 from Charles II. The original members included Christopher Wren, Robert Boyle, John Wilkins, Sir Robert Moray, and William, Viscount Brouncker, who subsequently became the first president of the Society [357, 1144, 1351].

on their front door. Dantzig answered the door to find Neyman holding papers in his hand and, as the door opened, Neyman began excitedly telling Dantzig that he "written an introduction to one of [Dantzig's] papers" [10, p. 301]. Dantzig had no idea as to what Neyman was referring, but Neyman explained. Rather than being homework, the equations that Dantzig had worked out were two famous unsolved problems in statistics, and the paper Neyman held was the solution to the first of those two problems.

A year later, the now-solved equations were formally put together as Dantzig's doctoral dissertation. In 1950, Dantzig received a letter from Abraham Wald that included proofs of a paper. Wald had solved the second of the two equations not knowing about Dantzig's solutions and when he submitted it for publication, a reviewer informed Wald about Dantzig's dissertation. Wald contacted Dantzig suggesting they publish the paper together. The first solution was published in 1940, "On the non-existence of tests of 'Student's' hypothesis having power functions independent of σ" by Dantzig [315] and the second solution was published in 1951 "On the fundamental lemma of Neyman and Pearson" by Dantzig and Wald [316].

G.B. Dantzig

George Bernard Dantzig went on to a distinguished career at Stanford University in the department of Operations Research, which he founded in 1966. In 1975 President Gerald Ford awarded Dantzig a National Medal of Science "for inventing Linear Programming and for discovering the Simplex Algorithm that led to wide-scale scientific and technical applications to important problems in logistics, scheduling, and network optimization, and to the use of computers in making efficient use of the mathematical theory" [287, 824]. George Bernard Dantzig died peacefully on 13 May 2005 at his home in Stanford, California, at the age of 90.

The earliest discussions of permutation methods appeared in the literature when Jerzy Spława-Neyman foreshadowed the use of permutation methods in a 1923 article "On the application of probability theory to agricultural experiments"; however, there is no indication that any of those who worked to establish the field of permutation methods were aware of the work by Spława-Neyman, which was not translated from its original Polish-language text until 1990 by D.M. Dabrowska and T.P. Speed [309]. In this early article, Spława-Neyman introduced a permutation model for the analysis of field experiments conducted for the purpose of comparing a number of crop varieties [1312]. The article was part of his doctoral thesis submitted to the University of Warsaw in 1924 and was based on research that he had previously carried out at the Agricultural Institute of Bydgoszcz in northern

Poland [1304]. A brief synopsis of the article by Spława-Neyman can be found in Scheffé [1231, p. 269, fn. 13]. Additionally, an introduction by Speed to the 1990 translation of "On the application of probability theory to agricultural experiments" by Dabrowska and Speed also provides a useful summary [1304], and a commentary on the translated article by D.B. Rubin is especially helpful in understanding the contribution made to permutation methods by Spława-Neyman in 1923 [1203]. See also a 1966 article by Stephen Fienberg and Judith Tanur in *International Statistical Review* [430].

Spława-Neyman introduced his model for the analysis of field experiments based on the completely randomized model, a model that Joan Fisher Box, R.A. Fisher's daughter, described as "a novel mathematical model for field experiments" [195, p. 263]. He described an urn model for determining the variety of seed each plot would receive. For m plots on which v varieties might be applied, there would be $n = m/v$ plots exposed to each variety. Rubin contended that this article represented "the first attempt to evaluate…the repeated-sampling properties of statistics over their non-null randomization distributions" [1203, p. 477] and concluded that the contribution was uniquely and distinctly Spława-Neyman's [1203, p. 479]. Rubin contrasted the contributions of Spława-Neyman and Fisher, which he observed, were completely different [1203, p. 478]. As Rubin summarized, Fisher posited a null hypothesis under which all values were known, calculated the value of a specified statistic under the null hypothesis for each possible permutation of the data, located the observed value in the permutation distribution, and calculated the proportion of possible values as or more unusual than the observed value to generate a probability value. In contrast, Spława-Neyman offered a more general plan for evaluating the proposed procedures [1203]. J.F. Box, commenting on the differences between Spława-Neyman and Fisher, noted that the conflict between Spława-Neyman and Fisher was primarily conditioned by their two different approaches: "Fisher was a research scientist using mathematical skills, Neyman a mathematician applying mathematical concepts to experimentation" [195, p. 265].[4]

2.2.2 Fisher and the Binomial Distribution

Ronald Aylmer Fisher was arguably the greatest statistician of any century [576, 738, 1483], although it is well known that his work in genetics was of comparable status, where geneticists know him for his part in the Wright–Fisher–Haldane theory of the neo-Darwinian synthesis, the integration of Darwinian natural selection with Mendelian genetics, and his 1930 publication of *The Genetical Theory of Natural*

[4] Fisher and Neyman differed in other ways as well. In general, they differed on the fundamental approach to statistical testing, with Fisher's ideas on significance testing and inductive inference and Neyman's views on hypothesis testing and inductive behavior; see an excellent summary in a 2004 article by Hubbard [663] as well as a comprehensive account of the controversy by Gigerenzer, Swijtink, Porter, and Daston published in 1989 [512, pp. 90–106].

Selection [80,576]. As L.J. Savage expressed it: "[e]ven today [1976], I occasionally meet geneticists who ask me whether it is true that the great geneticist R. A. Fisher was also an important statistician" [401, 1226, p. 445].

R.A. Fisher

Ronald Aylmer Fisher held two chairs in genetics, but was never a professor of statistics. Fisher was born on 17 February 1890 and even as a youth his eyesight was very poor; therefore, he was forbidden by his doctors to work by electric light [1477]. For example, James F. Crow, of the Genetics Department at the University of Wisconsin, recalled his first meeting with Fisher at North Carolina State University at Raleigh: "I . . . realized for the first time that in poor light Fisher was nearly blind" [297, p. 210]. Studying in the dark gave Fisher exceptional ability to solve mathematical problems entirely in his head, and also a strong geometrical sense [1477]. Fisher was educated at the Harrow School and the University of Cambridge [628]. His undergraduate degree was in mathematics at Gonville & Caius College, University of Cambridge, (informally known as Cambridge University or, simply, Cambridge), where he graduated as a Wrangler in 1912.[5]

After graduation, Fisher spent a post-graduate year studying quantum theory and statistical mechanics under mathematician and physicist James Hopwood Jeans and the theory of errors (i.e., the normal distribution) under astronomer and physicist Frederick John Marrian Stratton. It should be mentioned that while at the University of Cambridge, Fisher took only a single course in statistics. After graduating from Cambridge, Fisher taught mathematics and physics in a series of secondary schools and devoted his intellectual energies almost exclusively to eugenics. As Stigler reported, between 1914 and 1920 Fisher published 95 separate pieces; 92 in eugenics, one in statistical genetics, and two in mathematical statistics [1323, p. 24].

In 1918, almost simultaneously, Fisher received two invitations: one for a temporary position as a statistical analyst at the Rothamsted Experimental Station and the second from Karl Pearson at the Galton Biometric Laboratory at University College, London. The position at the Galton Biometric Laboratory came with the condition that Fisher teach and publish only what Pearson approved [778, p. 1020]; consequently, in 1919 Fisher took the position at the Rothamsted Experimental Station. As George Box described it:

(continued)

[5] Those students doing best on the examinations were designated as "Wranglers." More specifically, the 40 top-scoring students out of the approximately 100 mathematics graduates each year were designated as Wranglers, whereas 400–450 students graduated from the University of Cambridge annually at that time. Wranglers were rank-ordered according to scores on their final mathematics examination, which was a 44-h test spread over 8 days [713, p. 657].

> Fisher rejected the security and prestige of working under Karl Pearson in the most distinguished statistical laboratory in Britain and at that time certainly in the world. Instead, he took up a temporary job as the sole statistician in a small agricultural station in the country [191, p. 792].

Fisher left Rothamsted in 1933 after 14 years to assume the position of Galton Professor of Eugenics at University College, London. This was an uncomfortable arrangement for Fisher, in that the Department of Applied Statistics at University College, London, founded by Karl Pearson, was split into two departments upon Karl Pearson's retirement in 1933: the Department of Applied Statistics with Karl Pearson's son Egon as the head, and the Department of Eugenics with Fisher as the head and Galton Professor of Eugenics. Consequently, Fisher was barred from teaching statistics [816, p. 2]. When World War II broke out in 1939, Fisher's Department of Eugenics was evacuated from London and the faculty dispersed. Fisher did not find another position until 1943 when he returned to the University of Cambridge as the Arthur Balfour Chair of Genetics, succeeding the geneticist R.C. Punnett [1477]. Fisher was elected Fellow of the Royal Society in 1929 and knighted by Queen Elizabeth II in 1952. Sir Ronald Aylmer Fisher F.R.S. died in Adelaide, Australia, following complications from surgery on 29 July 1962 at the age of 72 [197, 814, 816, 1497, pp. 420–421].

Although Fisher published a great deal, his writing style sometimes confounded readers. There are numerous stories about the obscurity of Fisher's writing. To put it bluntly, Fisher did not always write with style and clarity. W.S. Gosset was once quoted as saying:

> [w]hen I come to Fisher's favourite sentence — "It is therefore obvious that ..." — I know I'm in for hard work till the early hours before I get to the next line (Gosset, quoted in Edwards and Bodmer [398, p. 29]).

Fisher's classical work on *The Genetical Theory of Natural Selection*, which has been described as the deepest book on evolution since Darwin's *On the Origin of Species* [398, p. 27], has come in for both considerable criticism and praise for his writing style. W.F. Bodmer stated:

> [m]any a terse paragraph in his classical work *The Genetical Theory of Natural Selection* has been the basis for a whole new field of experimental and theoretical analysis (Bodmer, quoted in Edwards and Bodmer [398, p. 29],

and Fred Hoyle, the English astronomer, once wrote:

> I would like to recommend especially R.A. Fisher's *The Genetical Theory of Natural Selection* for its brilliant obscurity. After two or three months of investigation it will be found possible to understand some of Fisher's sentences (Hoyle, quoted in Edwards and Bodmer [398, p. 29]).

Fisher's 1925 textbook *Statistical Methods for Research Workers* has also come under fire for its difficulty. M.G. Kendall has been quoted as saying:

> [s]omebody once said that no student should attempt to read [*Statistical Methods for Research Workers*] unless he had read it before (Kendall, quoted in Edwards and Bodmer [398, p. 29]).

While chemistry had its Mendeleev, mathematics its Gauss, physics its Einstein, and biology its Darwin, statistics had its Fisher. None of these scientists did all the work, but they did the most work, and they did it more eloquently than others. When simplifying history it is tempting to give each of these scientists too much credit as they did the important work in building the foundation on which to develop future works. On the other hand, the contributions of R.A. Fisher to the field of statistics cannot be overstated. There are few achievements in the history of statistics to compare—in number, impact, or scope—with Fisher's output of books and papers. In fact, Fisher was not trained as a statistician; he was a Cambridge-trained mathematician, with an extraordinary command of special functions, combinatorics, and n-dimensional geometry [1226].

In 1952, when presenting Fisher for the Honorary degree of Doctor of Science at the University of Chicago, W. Allen Wallis described Fisher in these words:

> [h]e has made contributions to many areas of science; among them are agronomy, anthropology, astronomy, bacteriology, botany, economics, forestry, meteorology, psychology, public health, and — above all — genetics, in which he is recognized as one of the leaders. Out of this varied scientific research and his skill in mathematics, he has evolved systematic principles for the interpretation of empirical data; and he has founded a science of experimental design. On the foundations he has laid down, there has been erected a structure of statistical techniques that are used whenever men attempt to learn about nature from experiment and observation (Wallis, quoted in Box [191, p. 791]).

In 1922 Fisher published a paper titled "On the mathematical foundations of theoretical statistics" that Stigler has called "the most influential article on…[theoretical statistics] in the twentieth century," describing the article as "an astonishing work" [1322, p. 32]. It is in this paper that the phrase "testing for significance" appears in print for the first time [816, p. 11]. However, as Bartlett explained in the first Fisher Memorial Lecture in 1965, while it is customary for statisticians to concentrate on Fisher's publications in statistics, his work in genetics was of comparable status [80, p. 395]. Fisher's interest in statistics began with a paper in 1912 [441] and his subsequent contributions can be divided into three main lines: exact sampling distribution problems, a general set of principles of statistical inference, and precise techniques of experimental design and analysis [80, p. 396]. In the present context, Fisher's contributions to permutation methods is the focus, especially his development of exact probability analysis.[6]

[6] The standard biography of R.A. Fisher is that written by his daughter in 1978, Joan Fisher Box [195], but others have provided more specialized biographies, including those by P.C. Mahalanobis [868], F. Yates [1474], F. Yates and K. Mather [1477], M.S. Bartlett [80], S.M. Stigler [1322,1323], C.R. Rao [1155], W.H. Kruskal [778], M.J.R. Healy [607], N.S. Hall [575], E.L. Lehmann [816],

"Student" and Sampling Distributions

In 1925 R.A. Fisher published his first book, titled *Statistical Methods for Research Workers* [448]. It was in this book that Fisher acknowledged that "[t]he study of the exact distributions of statistics commences in 1908 with 'Student's' paper *The Probable Error of a Mean*" [448, p. 23]. In neither of Student's 1908 papers, "The probable error of a mean" [1331] or "The probable error of a correlation coefficient" [1330] does Student make any reference to a previous use of the method and Egon Pearson stated in 1939 that Student's 1908 paper was the first instance of the use of exact distributions that was known to him [1094, p. 223].

The story of Student and the problem of finding the distribution of the standard deviation and the ratio of the mean to the standard deviation (the t statistic) is common knowledge. "Student" was born, as is well known, William Sealy Gosset on 13 June 1876 in Canterbury, England. He attended Winchester College and New College, University of Oxford (informally known as Oxford University or, simply, Oxford), graduating in 1899 with degrees in mathematics and chemistry. That same year he joined the Dublin Brewery of Messrs. Arthur Guinness Son & Company, Ltd. at St. James' Gate. In 1906–1907 Student was on leave from Guinness for a year's specialized study on probability theory. He spent the greater part of the year working at or in close contact with Karl Pearson's Biometric Laboratory at University College, London, where he first tackled the problem of inference from small samples empirically through a sampling experiment [177].

Student used as his study population a series of 3,000 pairs of measurements that had been published in an article on criminal anthropometry by William Robert Macdonell in *Biometrika* in 1902 [862]. The data consisted of measurements obtained by Macdonell of the height and length of the left middle finger of 3,000 criminals over 20 years of age and serving sentences in the chief prisons of England and Wales [862, p. 216]. (Student [1331, p. 13] lists page 219 for the Macdonell data, but the data used actually appear on page 216.) For the sampling experiment, Student recorded the data on 3,000 pieces of cardboard that were constantly shuffled and a card drawn at random, resulting in the 3,000 paired measurements arranged in random order. Then, each consecutive set of four measurements was selected as a sample—750 in all—and the mean, standard deviation, and correlation of each sample was calculated [see 1344]. He plotted the empirical distributions of the statistics and compared them to the theoretical ones he had derived. Using chi-squared

(continued)

L.J. Savage [1226], and G.E.P. Box [191]. The collected papers of R.A. Fisher are posted at http://www.adelaide.edu.au/library/special/digital/fisherj/. In addition, two large volumes of the selected correspondence of R.A Fisher were published in 1983 and 1990 by J.H. Bennett [96, 97].

tests for goodness of fit between the empirical and theoretical distributions, Student deemed the results to be satisfactory, noting "if the distribution is approximately normal our theory gives us a satisfactory measure of the certainty to be derived from a small sample" [1331, p. 19].

Egon Pearson had this to say of the 1908 paper of Student on small samples:

> [i]t is probably true to say that this investigation published in 1908 has done more than any other single paper to bring these subjects within the range of statistical inquiry; as it stands it has provided an essential tool for the practical worker, while on the theoretical side it has proved to contain the seed of new ideas which have since grown and multiplied an hundredfold [1094, p. 224].

During his 30 years of scientific activity, Student published all of his work under the pseudonym "Student" with only one exception, when reading a paper before the Industrial and Agricultural Research Section of the Royal Statistical Society in the Spring of 1936 [1034]. The reason for the pseudonym was a policy by Guinness against work done for the firm being made public. Allowing Gosset to publish under a pseudonym was a concession by Guinness that resulted in the birth of the statistician "Student" [813]. William Sealy Gosset died on 16 October 1937 at the age of 61 while still employed at Guinness.

In 1925, 2 years after Spława-Neyman introduced a permutation model for the analysis of field experiments, Fisher calculated an exact probability value using the binomial probability distribution in his first book: *Statistical Methods for Research Workers* [448, Sect. 18]. Although the use of the binomial distribution to obtain a probability value is not usually considered to be a permutation test per se, Scheffé considered it the first application in the literature of a permutation test [1230, p. 318]. Also, the binomial distribution does yield an exact probability value and Fisher found it useful in calculating the exact expected values for experimental data. Fisher wrote that the utility of any statistic depends on the original distribution and "appropriate and exact methods," which he noted have been worked out for only a few cases. He explained that the application is greatly extended as many statistics tend to the normal distribution as the sample size increases, acknowledging that it is therefore customary to assume normality and to limit consideration of statistical variability to calculations of the standard error or probable error.[7] That said, in

[7] Early on, the probable error was an important concept in statistical analysis and was defined as one-half the interquartile range. In terms of the normal distribution, the probable error is 0.6745 times the standard error. Therefore, as a test of significance a deviation of three times the probable error is effectively equivalent to one of twice the standard error [292, 448, pp. 47–48]. "Probable error" instead of "standard error" was still being used in the English-speaking countries in the 1920s and far into the 1930s; however, "probable error" was rarely used in Scandinavia or in the German-speaking countries [859, p. 214].

Table 2.1 Weldon's data on dice cast 26,306 times with a face showing five or six pips considered a success

Number of dice with a 5 or a 6	Observed frequency	Expected frequency	Difference frequency
0	185	202.75	−17.75
1	1,149	1,216.50	−67.50
2	3,265	3,345.37	−80.37
3	5,475	5,575.61	−100.61
4	6,114	6,272.56	−158.56
5	5,194	5,018.05	+175.95
6	3,067	2,927.20	+139.80
7	1,331	1,254.51	+76.49
8	403	392.04	+10.96
9	105	87.12	+17.88
10	14	13.07	+0.93
11	4	1.19	+2.81
12	0	0.05	−0.05
Total	26,306	26,306	−0.02

Chap. III, Sect. 18 of *Statistical Methods for Research Workers*, Fisher considered the binomial distribution and provided two examples.

The first example utilized data from the evolutionary biologist Walter Frank Raphael Weldon. Weldon threw 12 dice 26,306 times for a total of 315,672 observations, recording the number of times a 5 or a 6 occurred. Fisher did not provide a reference for the Weldon data, but the source was a letter from Weldon to Francis Galton dated 2 February 1894 in which Weldon enclosed the data for all 26,306 throws and asked Galton his opinion as to the validity of the data [717, pp. 216–217]. Fisher used the binomial distribution to obtain the exact expected value for each of the possible outcomes of 0, 1, ..., 12. For example, the binomial probability for six of 12 dice showing either a 5 or a 6 is given as

$$p(6|12) = \binom{12}{6}\left(\frac{2}{6}\right)^{6}\left(\frac{4}{6}\right)^{12-6} = (924)(0.0014)(0.0878) = 0.1113\ .$$

Multiplying 0.1113 by $n = 26{,}306$ gives an expectation of 2,927.20. Table 2.1 summarizes the Weldon dice data; see also Fisher [448, p. 67] and Pearson [1107, p. 167]. Fisher concluded the dice example by calculating a chi-squared goodness-of-fit test and a normal approximation to the discrete binomial distribution.

For the second example, Fisher analyzed data from Arthur Geissler on the sex ratio at birth in German families. Here again, Fisher did not provide a reference to the Geissler data, but it was taken from the sex-ratio data obtained by Geissler from hospital records in Saxony and published in *Zeitschrift des Königlich Sächsischen Statistischen Bureaus* in 1889 [504]. The data consisted of the number of males in 53,680 families, ranging from 0 to 8 males. Geissler's estimate of the sex ratio for

Table 2.2 Geissler's data on the sex ratio in German families with expected values and differences, and Fisher's expected values and differences

Geissler's data and expected values				Fisher's expected values	
Number of males	Observed sibships	Expected sibships	Difference (Obs – Exp)	Expected sibships	Difference (Obs – Exp)
8	342	264.64	+77.36	264.30	+77.70
7	2,092	1,995.88	+96.12	1,993.78	+98.22
6	6,678	6,584.71	+93.29	6,580.24	+97.76
5	11,929	12.413.82	−484.82	12,409.87	−480.87
4	14,959	14,626.99	+332.01	14,627.60	+331.40
3	10,649	11,030.22	−381.22	11,034.65	−385.65
2	5,331	5,198.69	+132.31	5,202.65	+128.35
1	1,485	1,400.08	+84.92	1,401.69	+83.31
0	215	164.96	+50.04	165.22	+49.78
Total	53,680	53,679.99	+0.01	53,680.00	0.00

the population in Saxony was obtained by simply calculating the mean proportion of males in his data. Table 2.2 summarizes the Geissler sex-ratio data [793, p. 154]. In this second example, Fisher never specified a value for p, but H.O. Lancaster, in a reanalysis of Geissler's data, gave the value as $p = 0.5147676$ [793], which translates to a sex ratio of 1.061.[8] Working backwards from Fisher's analysis, it is apparent that he used $p = 0.5146772$. Thus, for example, the binomial probability for five males is actually given by

$$p(5|8) = \binom{8}{5}(0.5146772)^5(0.4853228)^{8-5} = (56)(0.0361)(0.1143) = 0.2312\ .$$

Multiplying 0.2312 by $n = 53{,}680$ gives an expectation of 12,409.87, which agrees with Fisher's expected value.

In both these early examples Fisher demonstrated a preference for exact solutions, eschewing the normal approximation to the discrete binomial distribution even though the sample sizes were very large. While exact binomial probability values are perhaps not to be considered as permutation tests, Fisher was to go on to develop many permutation methods and this early work provides a glimpse into how Fisher advanced exact solutions for statistical problems.

2.2.3 Geary and Correlation

In 1927, R.C. Geary was the first to use an exact analysis to demonstrate the utility of asymptotic approaches for data analysis in an investigation of the properties of correlation and regression in finite populations [500].

[8] For comparison, the sex ratio at birth in Germany in 2013 was 1.055.

R.C. Geary

Robert Charles (Roy) Geary was a renowned Irish economist and statistician who earned his B.Sc. degree from University College, Dublin, in 1916 and pursued graduate work at the Sorbonne in Paris where he studied under Henri Lebesgue, Émile Borel, Élie Cartan, and Paul Langevin [1307]. Geary's early contributions in statistics were greatly influenced by the work of R.A. Fisher, although in later years Geary's attention turned towards more social issues, e.g., poverty and inequality [1306]. Geary did work on permutation tests early in his career and was an early critic of reliance on the normal distribution. In 1947, for example, he considered the problem of statistics and normal theory, calling for future statistics textbooks to include the phrase, "Normality is a myth; there never was, and never will be, a normal distribution" [501, p. 241].

Geary founded the Central Statistics Office of Ireland in 1960 and the Economic Research Institute (later, the Economic and Social Research Institute) in 1949, and was head of the National Accounts Branch of the United Nations from 1957 to 1960. Interestingly, more than half of Geary's 127 publications were written in the 1960s after Geary had reached 65 years of age. Robert Charles Geary retired in 1966 and passed away on 8 February 1983 at the age of 86 [1305, 1306].

In 1927 Geary devoted a paper to "an examination of the mathematical principles underlying a method for indicating the correlation . . . between two variates," arguing that "the formal theory of correlation . . . makes too great demands upon the slender mathematical equipment of even the intelligent public" [500, p. 83]. Geary provided a number of example analyses noting "[w]e are not dealing with a sample drawn from a larger universe" [500, p. 87] and addressed the problem of deciding significance when calculating from a known limited universe. One example that Geary provided was based on the assertion that cancer may be caused by the over consumption of "animal food." Geary investigated the ways that cancer mortality rates varied with the consumption of potatoes in Ireland, drawing up a contingency table showing 151 poor-law unions in Ireland arranged according to their percentage of deaths from cancer during the years 1901–1910 and the acreage of potatoes per 100 total population.[9] Table 2.3 summarizes Geary's data on cancer and potato consumption [500, p. 94].

In this investigation, Geary considered potato consumption and the incidence of cancer deaths in Ireland. Geary categorized each of the 151 poor law unions

[9] The Irish Poor Law of 1838 was an attempt to ameliorate some of the problems arising out of widespread poverty in the early 1800s in Ireland. Influenced by the Great Reform Act of 1834 in England (q.v. page 11), Ireland was originally divided into 131 poor law unions, each with a workhouse at its center.

Table 2.3 Percentage of deaths from cancer to all deaths during the 10 years 1901–1910 cross classified by acreage of potatoes per 100 total population

Cancer deaths as percentage of total deaths 1901–1910	Number of poor law unions in which acreage of potatoes per 100 persons in 1911 was			
	Under 15.5	15.5–20.5	Over 20.5	Number of unions
Under 3.5 %	12	24	12	48
3.5–4.5 %	18	14	16	48
Over 4.5 %	20	17	18	55
Number of unions	50	55	46	151

as a percentage of cancer deaths to overall deaths in the union; cancer deaths less than 3.5 % of total deaths (48 poor law unions), cancer deaths 3.5–4.5 % of total deaths (48 poor law unions) and cancer deaths greater than 4.5 % of total deaths (55 poor law unions). Table 2.3 illustrates the marginal distribution of 48, 48, and 55 poor law unions. He repeated the experiment holding the marginal frequency totals constant, and found that cell arrangements greater than those of the actual experiment occurred in 231 of 1,000 repetitions, concluding that the relationship between potato consumption and cancer was not statistically significant.

2.3 Fisher and the Variance-Ratio Statistic

Because of its importance, some historical perspective on Fisher's variance-ratio z test and the analysis of variance is appropriate. Fisher's variance-ratio z test statistic is given by

$$z = \frac{1}{2} \log_e \left(\frac{v_1}{v_0} \right) , \tag{2.1}$$

where $v_1 = MS_{\text{Between}} = MS_{\text{Treatment}}$ and $v_0 = MS_{\text{Within}} = MS_{\text{Error}}$ in modern notation, and which Fisher termed, for obvious reasons, the "variance-ratio" statistic. In a 1921 article on grain yields from Broadbalk wheat from the Rothamsted Experimental Station (q.v. page 57) in *The Journal of Agricultural Science*, Fisher partitioned the total sum of squares of deviations from the mean into a number of independent components and made estimates of the component variances by associating each sum of squares with its appropriate degrees of freedom [445]. Fisher made the analysis of variance source table explicit in 1923 in a second article on "Studies in crop variation II," subtitled "The manurial response of different potato varieties," in *The Journal of Agricultural Science* with his assistant

Winifred A. Mackenzie [462].[10,11] The analysis of variance appears in this article with Mackenzie for the first time in its entirety, although it is not reflected in the title [191, p. 795].[12] Experimental randomization is also firmly established in this article.[13] After the algebraic identity between the total sum of squares and the within- and between-treatments sum of squares had been presented, Fisher and Mackenzie stated:

> [i]f all the plots were undifferentiated, as if the numbers had been mixed up and written down in random order, the average value of each of the two parts is proportional to the number of degrees of freedom in the variation of which it is compared [462, p. 315], quoted in [191, p. 795].

However, as Joan Fisher Box explained, the analysis was incorrect because the trial was actually a split-plot design as it incorporated a third factor: potassium. At the time of the writing of the article, 1923, Fisher did not fully understand the rules of the analysis of variance, nor the role of randomization [261]. Fisher quickly corrected this in the first edition of *Statistical Methods for Research Workers* published in 1925 [448, p. 238].

In *Statistical Methods for Research Workers* Fisher detailed the analysis of variance in Chap. VII on "Intraclass correlations and the analysis of variance" [448]. An important observation by J.F. Box, is that it tends to be forgotten that prior to 1920, problems that would later be dealt with by the analysis of variance were thought of as problems in correlation [195, p. 100]; thus, R.A. Fisher introduced the subject of analysis of variance in terms of its relation to the intraclass correlation coefficient. The relationship between the intraclass correlation coefficient, r_I, and Fisher's z is given by

$$z = \frac{1}{2}\log_e\left\{\left(\frac{k}{k-1}\right)\left[\frac{1+r_I(n-1)}{1-r_I}\right]\right\} ,$$

where n is the number of observations in each of k treatments.

By way of example, consider two samples of n_1 and n_2 observations, each sample drawn from one of two populations consisting of normally distributed variates with

[10] Mackenzie is sometimes spelled "Mackenzie" [195] and other times "MacKenzie" [191, 576, 720]. In the original article, Mackenzie is all in upper-case letters.

[11] The experiment on potatoes had been conducted by Thomas Eden at the Rothamsted Experimental Station, wherein each of twelve varieties of potatoes had been treated with six different combinations of manure [191].

[12] Previously, in 1918 in an article on Mendelian inheritance in *Eugenics Review*, Fisher had coined the term "analysis of variance" [443]; see also a 2012 article by Edwards and Bodmer on this topic [398, p. 29].

[13] This 1923 article by Fisher and Mackenzie is often cited as the first randomized trial experiment [484, 517, 893, 925]. However, the first documented publication of a randomized trial experiment was by the American philosopher Charles Sanders Peirce and his colleague at Johns Hopkins University, Joseph Jastrow, in 1885 [1113]; see also, in this regard, discussions by Neuhauser and Diaz [1030, pp. 192–195], Stigler [1321], and an autobiography by Jastrow [682].

equal population variances. It can be shown that the distribution of z approaches normality as $\min(n_1, n_2) \to \infty$, with mean and variance given by

$$\bar{z} = \frac{1}{2}\left(\frac{1}{n_2 - 1} - \frac{1}{n_1 - 1}\right)$$

and

$$s_z^2 = \frac{1}{2}\left(\frac{1}{n_2 - 1} + \frac{1}{n_1 - 1}\right) ,$$

respectively [36, p. 439]. These results stimulated Fisher to prefer the designation z for the analysis of variance test statistic over the F proposed by Snedecor in 1934 [1289].

2.3.1 Snedecor and the *F* Distribution

G.W. Snedecor was the director of the Statistical Laboratory at Iowa State College (technically, Iowa Agricultural College and Model Farm) and was instrumental in introducing R.A. Fisher and his statistical methods to American researchers.

G.W. Snedecor

George Waddle Snedecor earned his B.S. degree in mathematics and physics from the University of Alabama in 1905 and his A.M. degree in physics from the University of Michigan in 1913, whereupon Snedecor accepted a position as Assistant Professor of mathematics at Iowa State College of Agriculture (now, Iowa State University). Snedecor's interest in statistics led him to offer the first course in statistics in 1915 on the *Mathematical Theory of Statistics* at Iowa State College of Agriculture. In 1933, Snedecor became the Director of the Statistical Laboratory, remaining there until 1947. Snedecor was responsible for inviting R.A. Fisher to Iowa State College during the summers of 1931 and 1936 to introduce statistical methods to faculty and research workers [295].

In 1937, Snedecor published a textbook on *Statistical Methods*, subtitled *Applied to Experiments in Agriculture and Biology*, which was a phenomenal success selling more than 200,000 copies in eight editions. The first five editions were authored by Snedecor alone and the next three editions were co-authored with William Gemmell Cochran. Snedecor's *Statistical Methods* roughly covered the same material as Fisher's *Statistical Methods for Research Workers*, but also included material from Fisher's book on *The Design of Experiments*, such as factorial experiments, randomized blocks,

(continued)

Latin squares and confounding [816, p. 27]. Joan Fisher Box wrote in her biography of her father that "[i]t was George W. Snedecor, working with agricultural applications, who was to act as midwife in delivering the new statistics in the United States" [195, p. 313]. George Waddle Snedecor died on 15 February 1974 at the age of 92 [59, 243, 611].

Fisher had, in the first edition of *Statistical Methods for Research Workers*, provided a brief tabulation of critical values for z—Table VI in [448]—corresponding to a 5 % level of significance, noting "I can only beg the reader's indulgence for the inadequacy of the present table" [448, p. 24]. In 1934, apparently in an attempt to eliminate the natural logarithms required for calculating z, Snedecor [1289] published tabled values in a small monograph for Fisher's variance-ratio z statistic and rechristened the statistic, F [1289, p. 15]. Snedecor's F-ratio statistic was comprised of

$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}} = \frac{MS_{\text{Treatment}}}{MS_{\text{Error}}} ,$$

whereas Fisher had used

$$z = \frac{1}{2}\log_e\left(\frac{v_1}{v_0}\right) = \frac{1}{2}\log_e(F) .$$

In terms of the intraclass correlation coefficient,

$$F = \left(\frac{k}{k-1}\right)\frac{1 + r_I(n-1)}{1 - r_I}$$

and, conversely,

$$r_I = \frac{(k-1)F - k}{(k-1)F + (n-1)k} .$$

It has often been reported that Fisher was displeased when the variance-ratio z statistic was renamed the F-ratio by Snedecor, presumably in honor of Fisher; see also discussions by Box [195, p. 325] and Hall on this topic [575, p. 295]. Fisher recounted in a letter to H.W. Heckstall-Smith in 1956 that "I think it was only an afterthought that led Snedecor to say that the capital letter F he had used was intended as a compliment to myself" [97, p. 319].[14] In this same letter, Fisher also wrote that he had added a short historical note in the 12th edition of *Statistical*

[14] H.W. Heckstall-Smith, Headmaster, Chippenham Grammar School, had written to Fisher requesting permission to quote from Fisher in an article he was preparing for a medical journal.

Methods for Research Workers published in 1954 that he "hoped [would] prevent expositors from representing the F-test ... with the z-test" [97, p. 319]. On this topic, in a 1938 letter to Snedecor, Fisher objected to the assignment of the symbol F to the variance-ratio z statistic, and used the letter to point out that P.C. Mahalanobis had previously published tabled values of the variance-ratio z statistic using a different symbol, although Snedecor apparently produced his F-ratio with no knowledge of the Mahalanobis tables [195, p. 325].

Indeed, in 1932 Mahalanobis, responding to complaints from field workers who were not familiar with the use of natural logarithms and had difficulty with Fisher's variance-ratio z statistic as given in Eq. (2.1), published six tables in *Indian Journal of Agricultural Science*. Two tables were designed for working with ordinary logarithms (base 10 instead of base e), two tables were designed for working directly with the ratio of standard deviations instead of variances, and two tables were designed for the ratio of variances without recourse to natural logarithms, with one table in each set corresponding to the 5 % level of significance and the other set to the 1 % level of significance [867]. Fisher avoided using the symbol F in *Statistical Tables for Biological, Agricultural and Medical Research* published with Yates in 1938, as Fisher felt that the tabulation of Mahalanobis had priority [195, p. 325].

2.4 Eden–Yates and Non-normal Data

In 1933 Frank Yates succeeded R.A. Fisher as head of the Statistics Department at the Rothamsted Experimental Station, a post he held for a quarter of a century.

F. Yates

Frank Yates graduated from St. John's College, University of Cambridge, with a B.A. degree in mathematics in 1924 and earned his D.Sc. in mathematics from Cambridge in 1938. His first important job was as research officer and mathematical advisor to the Geodetic Survey of the Gold Coast (presently, Ghana). In August 1931, Yates joined Fisher at the Rothamsted Experimental Station as an Assistant Statistician. Within 2 years, Fisher had left Rothamsted and Yates became head of the Statistics Department, a post which he held for 25 years until 1958. From 1958 until his retirement in 1968, Yates was Deputy Director of Rothamsted [437]. Although retired, Yates maintained an office at Rothamsted as an "Honorary Scientist" in the Computing Department and all told, was at Rothamsted for a total of 60 years. Perhaps Frank Yates' greatest contribution to statistics was his embrace of the use of computing to

(continued)

The article, with M.G. Ellis, was eventually published in the journal *Tubercle* in December of 1955 under the title "Fun with statistics" [409].

solve statistical problems [633, p. 4]. In 1948 Yates was elected Fellow of the Royal Society. Frank Yates F.R.S. passed away on 17 June 1994 at the age of 92 [369, 436, 605, 606, 1028].

T. Eden

Little is known about Thomas Eden, except that he was at the Rothamsted Experimental Station as a crop ecologist in the Field Experiments Department from 1921 to 1927 and published several papers with Fisher on experimental design [377,378]. Upon leaving Rothamsted, Eden was employed as a chemist at the Tea Research Institute of Ceylon [575, p. 318]. Eden published a number of books in his lifetime, including *Soil Erosion* in 1933 [374], *Elements of Tropical Soil Science* in 1947 [375], and *Tea* in 1958 [376].

Like Geary in 1927 [500], Thomas Eden and Frank Yates utilized permutation methods in 1933 to compare a theoretical distribution to an empirical distribution [379]. Eden and Yates questioned the use of Fisher's variance-ratio z test in applications to non-normal data. Citing articles by Shewhart and Winters [1262] and Pearson and Adyanthāya [1100] in which small samples from non-normal and skewed populations had been investigated, Eden and Yates declared the results "inconclusive" [379, p. 7], despite an affirmation by "Student" that " 'Student's' distribution will be found to be very little affected by the sort of small departures from normality which obtain in most biological and experimental work" [1332, p. 93] and Fisher's contention that he had "never known difficulty to arise in biological work from imperfect normality of variation" [440, p. 267]. Eden and Yates noted that from the perspective of the investigator who is using statistics as a tool "the theoretical distributions from which the samples were drawn bear no relationship to those he is likely to encounter" [379, p. 7] and listed three conditions which must be observed to compare a theoretical distribution with an empirical distribution:

1. Samples must be taken from one or more actual distribution(s).
2. The experimental procedure must correspond with what would be used on actual investigational data.
3. The departure of the distribution of the statistical tests from expectation must itself be tested for significance, and the sampling must be sufficiently extensive to give reliable evidence of the distribution in the neighborhood of the 0.05 and 0.01 levels of significance.

Some Historical Perspective

A little historical background will shed some light on the exchange between Fisher and Eden and Yates. In 1929, in the 8 June issue of *Nature*, Egon Pearson reviewed the second edition of Fisher's *Statistical Methods for Research Workers* that had been published in 1928. In that review, Pearson criticized Fisher's approach, noting:

> [a] large number of the tests developed are based upon the assumption that the population sampled is of the 'normal' form. . . . It does not appear reasonable to lay stress on the 'exactness' of tests, when no means whatever are given of appreciating how rapidly they become inexact as the population diverges from normality [1099, p. 867].

Fisher was deeply offended and he wrote a blistering reply to *Nature* that has not been preserved [816, p. 23]. Eventually, Fisher asked W.S. Gosset to reply for him, which Gosset did under his pseudonym "Student" in *Nature* on 20 July 1929, stating:

> [p]ersonally, I have always believed . . . that in point of fact 'Student's distribution will be found to be very little affected by the sort of small departures from normality which obtain in most biological and experimental work, and recent work on small samples confirms me in this belief. We should all of us, however, be grateful to Dr. Fisher if he would show us elsewhere on theoretical grounds what *sort* of modification of his tables we require to make when the samples with which we are working are drawn from populations which are neither symmetrical nor mesokurtic [1332, p. 93].

This was followed by a letter in *Nature* by Fisher on 17 August 1929, in which he rejected Gosset's suggestion that he should give some guidance on how to modify the t test for data from non-normal populations [440]. However, he did hint in this letter at the possibility of developing distribution-free tests. Finally, a rejoinder by E.S. Pearson appeared in *Nature* on 19 October 1929 [1092].

In hindsight, E.S. Pearson was probably correct in questioning the t test established by "Student" and proved by Fisher under the assumption of normality. Interestingly, the same argument also holds for the Neyman–Pearson statistical approach that requires the use of conjectured theoretical distributions such as the normal and gamma distributions. On a related note, Fisher seemed to have eventually accepted Pearson's normality concern since he introduced the notion of an exact permutation test a short time later.

In 1933 Eden and Yates observed that if evidence could be adduced showing that the distribution of z for treatments versus residuals was statistically identical to that expected from normal data, then the variance-ratio z statistic could be used with confidence when establishing significance to data of this type. Eden and Yates went on to examine height measurements of Yeoman II wheat shoots grown in eight blocks, each consisting of four sub-blocks of eight plots.[15] For the experiment, the observations were collapsed into four treatments randomly applied to four sub-blocks in each block. Thus, the experimental data consisted of $g = 4$ treatment groups and $b = 8$ treatment blocks for a total of

$$(g!)^{b-1} = (4!)^{8-1} = 4{,}586{,}471{,}424$$

possible arrangements of the observed data.[16] Eden and Yates chose a sample of 1,000 of these arrangements at random (now termed resampling) and generated a table listing the simulated probability values generated by the random sample and the theoretical counterparts to those probability values based on the normality assumption.[17]

Eden and Yates were able to reduce the considerable computations of the analysis by introducing "certain modifications" [379, p. 11]. Specifically, they observed that the block sum of squares and the total sum of squares would be constant for all 1,000 samples; consequently, the value of z for each sample would be uniquely defined by the value for the treatment sum of squares. This observation became increasingly valuable in later decades as researchers developed permutation versions of other statistical tests and increased the speed of computing by ignoring the components of equations that are invariant over permutation.

The simulated and theoretical probability values based on the normality assumption were compared by a chi-squared goodness-of-fit test and were found to be in close agreement, supporting the assumption of normality [379]. Eden and Yates therefore contended that Fisher's variance-ratio z statistic could be applied to data of this type with confidence. Specifically, Eden and Yates concluded:

> [t]he results of this investigation, which deals with an actual experimental distribution of a definitely skew nature and with a population extending over a wide range of values, show that in actual practice there is little to fear in the employment of the analysis of variance and the z test to data of a similar type [379, p. 16].

In 1935 Yates had one more opportunity to comment on this experiment, emphasizing once again reliance on the information contained in the sample alone. On March 28th, 1935, Neyman presented a paper before the Industrial and Agricultural Research Section of the Royal Statistical Society, later published in *Supplement to the Journal of the Royal Statistical Society* [1033], where Yates

[15] Yeoman wheat is a hybrid variety that resists wheat rust. It was developed and released in 1916 by Sir Rowland Biffen, Director of the Plant Breeding Institute at the University of Cambridge School of Agriculture.

[16] Because it is possible to hold one block constant and to randomize the remaining blocks with respect to the fixed block, it is only necessary to randomize $b-1$ blocks, thereby greatly decreasing the total number of possible arrangements. In this case, $(4!)^7 = 4{,}586{,}471{,}424$ instead of $(4!)^8 = 110{,}075{,}314{,}176$ randomizations.

[17] H.A. David has written that the 1933 Eden–Yates paper "may be regarded as introducing randomization [permutation] theory" [326, p. 70].

was a discussant. Referring back to the Yeoman II wheat shoot experiment, Yates commented:

> [w]hat the experiment does show is that the randomisation process effectively generates the distribution of z, and the need for the postulation of any parent population from which the thirty-two values are to be regarded as a sample is entirely avoided [1473, p. 165].

2.5 Fisher and 2 × 2 Contingency Tables

On 18 December 1934, R.A. Fisher (q.v. page 25) presented a paper describing the logic of permutation tests to the Royal Statistical Society, a paper that appeared in *Journal of the Royal Statistical Society* the following year [452].[18] Fisher did not expressly discuss permutation tests, but instead used the product of two binomial distributions to arrive at an exact probability value for a 2 × 2 contingency table. Here, Fisher described data on criminal same-sex twins from a study originally conducted by Lange [801, pp. 41–45]. Dr. Johannes Lange was Chief Physician at the Munich–Schwabing Hospital and Department Director of the German Experimental Station for Psychiatry (Kaiser Wilhelm Institute) in Munich. Lange had access to data on 37 pairs of criminal same-sex twins, including 15 monozygotic (identical) and 22 dizygotic (fraternal) twins, but in two cases of the monozygotic twins and five of the dizygotic twins, neither twin had been convicted, thus reducing the overall number of twin pairs to 30.

The data analyzed by Fisher consisted of 13 pairs of monozygotic twins and 17 pairs of dizygotic twins. For each of the 30 pairs of twins, one twin was known to be a convict. The study considered whether the twin brother of the known convict was himself "convicted" or "not convicted." Fisher observed that in 10 of the 13 cases of monozygotic twins, the twin brother was convicted, while in the remaining three cases, the twin was not convicted. Among the 17 pairs of dizygotic twins, two of the twins were convicted and 15 of the twins were not convicted. The data from Lange are summarized in Table 2.4. Fisher considered the many methods available for the analysis of a 2 × 2 table and suggested a new method based on the concept of ancillary information [816, p. 48–49]. Fisher explained: [i]f one blocked out the cell frequencies of Table 2.4 leaving only the marginal frequency totals, which provide no information by themselves, then the information supplied

[18] As was customary in scientific societies at the time, these special research papers were printed in advance and circulated to the membership of the society. Then, only a brief introduction was made by the author at the meeting and the remaining time was devoted to discussion. By tradition, the "proposer of the vote of thanks" said what was he thought was good about the paper, and the seconder said what he thought was not so good. Subsequently, there was a general discussion by the Fellows of the Society and often a number of prominent statisticians offered comments, suggestions, or criticisms [192, p. 41]. In this instance the discussants were Arthur Bowley, Leon Isserlis, Joseph Irwin, Julius Wolf, Egon Pearson, Major Greenwood, Harold Jeffreys, Maurice Bartlett, and Jerzy Neyman. As might be evident from the list of names, not all comments were constructive.

Table 2.4 Convictions of like-sex twins of criminals

Twin type	Convicted	Not convicted	Total
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

by the marginal frequency totals is "wholly ancillary" [452, p. 48].[19] Fisher was then concerned with the number of different ways the four cell frequencies could be filled, subject to the fixed marginal frequency totals. For these data, the maximum value of the convicted dizygotic cell is the minimum of the corresponding marginal frequency totals, and the minimum value of the convicted dizygotic cell is the greater of zero and the sum of the corresponding marginal frequency totals minus the total sample size. Thus, the number of possible configurations of cell frequencies completely specified by the number of dizygotic convicts is 13, ranging from 0, given by $\max(0, 17 + 12 - 30) = 0$, to 12, given by $\min(12, 17) = 12$.

The approach is clever and deserves consideration. Fisher posited that if the probability of a twin brother of a convict of monozygotic origin is denoted by p, then the probability that of 13 monozygotic twins $12-x$ have been convicted, while $x + 1$ monozygotic twins have escaped conviction, is given by the binomial

$$\frac{13!}{(12-x)!\,(1+x)!}\,p^{12-x}(1-p)^{1+x}\ .$$

The probability of the brother of a criminal known to be dizygotic being convicted is also p and the probability that 17 of these x have been convicted and $(17 - x)$ have never been convicted, is given by the binomial

$$\frac{17!}{x!\,(17-x)!}\,p^{x}(1-p)^{17-x}\ .$$

The probability of the simultaneous occurrence of the two events, given by the product of the respective probabilities, is therefore

$$\frac{13!\;17!}{(12-x)!\,(1+x)!\,x!\,(17-x)!}\,p^{12}(1-p)^{18}\ .$$

Fisher noted that the probability of any value of x occurring is proportional to

$$\frac{1}{(12-x)!\,(1+x)!\,x!\,(17-x)!}\ ,$$

[19] According to Lehmann [816, p. 48, fn. 1], this statement is in fact not completely true, although very nearly so. See also a 1977 article by Plackett in this regard [1137].

and on summing the series obtained over x, the absolute probability values are found to be

$$\frac{13!\ 17!\ 12!\ 18!}{30!} \times \frac{1}{(12-x)!\ (1+x)!\ x!\ (17-x)!}$$

[452, p. 49]. Thus, it is only necessary to compute the probability of one of the four cells; Fisher chose the dizygotic convicts, the lower-left cell in Table 2.4 with a frequency of 2. Computing the discrepancies from proportionality as great or greater than the observed configuration in Table 2.4, subject to the conditions specified by the ancillary information, yields for 2, 1, and 0 dizygotic convicts, a one-tailed probability of

$$\begin{aligned}
&P\{2|17,12,30\} + P\{1|17,12,30\} + P\{0|17,12,30\} \\
&\qquad = \frac{13!\ 17!\ 12!\ 18!}{30!\ 10!\ 3!\ 2!\ 15!} + \frac{13!\ 17!\ 12!\ 18!}{30!\ 11!\ 2!\ 1!\ 16!} + \frac{13!\ 17!\ 12!\ 18!}{30!\ 12!\ 1!\ 0!\ 17!} \\
&\qquad = 0.000449699 + 0.000015331 + 0.000000150\,,
\end{aligned}$$

which sums to approximately 0.0005.

The point of the twin example—that for small samples exact tests are possible, thereby eliminating the need for estimation—indicates an early understanding of the superiority of exact probability values computed from known discrete distributions over approximations based on assumed theoretical distributions. As Fisher pointed out, "[t]he test of significance is therefore direct, and exact for small samples. No process of estimation is involved" [451, p. 50]. In this regard, see also the fifth edition of *Statistical Methods for Research Workers* published in 1934 where Fisher added a small section on "The exact treatment of a 2 × 2 table" [450, Sect. 21.02]. The exact binomial solution proposed by Fisher was not without controversy [1197]. Indeed, Stephen Senn observed in 2012 that "statisticians have caused the destruction of whole forests to provide paper to print their disputes regarding the analysis of 2 × 2 tables" [1251, p. 33].

2.6 Yates and the Chi-Squared Test for Small Samples

In 1934 Frank Yates (q.v. page 37) published an article on contingency tables involving small frequencies and the chi-squared (χ^2) test of independence in *Supplement to the Journal of the Royal Statistical Society* [1472]. The stated purpose of the article was twofold: first, to introduce statisticians to Fisher's exact probability test, which was very new at the time, and to use Fisher's exact probability test as a gold standard against which the small-sample performance of the Pearson chi-squared test might be judged; and second, present the correction for continuity to the chi-squared test of independence, resulting in a better approximation to Fisher's exact probability test [633]. Yates motivated the discussion by asserting:

> [t]he χ^2 test is admittedly approximate, for in order to establish the test it is necessary to regard each cell value as normally distributed with a variance equal to the expected value, the whole set of values being subject to certain restrictions. The accuracy of this approximation depends on the numbers in the various cells, and in practice it has been customary to regard χ^2 as sufficiently accurate if no cell has an expectancy of less than 5 [1472, p. 217].[20]

The 1934 article by Yates soon became elevated to a classic as it introduced Yates' correction for continuity to chi-squared for 2 × 2 contingency tables. However, the article contained much more than the continuity correction for 2 × 2 contingency tables. In this 1934 article Yates referred to Fisher's calculation of the exact probability of any observed set of values in a 2 × 2 contingency table with given marginal frequency totals and compared chi-squared probability values, with and without the correction for continuity, with exact probability values for small 2 × 2 contingency tables. Yates used the exact probability values obtained from the discrete hypergeometric probability distribution to evaluate the corresponding probability values obtained from the continuous chi-squared distribution. It is notable that Yates referred to the exact probability values as the "true" probability values [1472, p. 222] and the exact probability values were used in this article as a benchmark against which to compare and validate the approximate probability values obtained from the chi-squared distribution.[21]

While there is much of importance in this classic paper, it is the generation of the exact probability values that is germane to a discussion of permutation methods. Although Yates only summarized the procedure by which he obtained the exact permutation values, the process is not difficult to reconstruct. Yates described the process:

> [i]n cases where N is not too large the distribution with any particular numerical values of the marginal totals can be computed quite quickly, using a table of factorials to determine some convenient term, and working out the rest of the distribution term by term, by simple multiplications and divisions. If a table of factorials is not available we may start with any convenient term as unity, and divide by the sum of the terms so obtained [1472, p. 219].

Note that N denotes the total number of observations. Here, in the last sentence of the quote, Yates identified a procedure that was to assume great importance in exact permutation methods; viz., probability values obtained from discrete distributions using recursion with an arbitrary initial value. The importance of this approach for the future of permutation methods should not be underestimated.

[20] As Hitchcock has noted, the variance equals the mean in the archetypical count model of the Poisson, and the normal approximates the Poisson when the mean is large [633, p. 2].

[21] It should be mentioned that because Yates was primarily interested in 2 × 2 contingency tables and, therefore, χ^2 was distributed as chi-squared with 1 degree of freedom, he obtained the requisite probability values from tables of the normal distribution since $\chi_1^2 = z^2$.

Fig. 2.1 Notation for a 2 × 2 contingency table as used by Yates [1472]

a	b	$N-n$
c	d	n
$N-n'$	n'	N

Next, Yates defined a 2 × 2 contingency table using the notation in Fig. 2.1, where $n \leq n' \leq \frac{1}{2}N$.

Giving due credit to Fisher, Yates showed that the probability value corresponding to any set of cell frequencies, a, b, c, d, was the hypergeometric point-probability value given by

$$\frac{n!\, n'!\, (N-n)!\, (N-n')!}{N!\, a!\, b!\, c!\, d!} \ .$$

Since the exact probability value of a 2 × 2 contingency table with fixed marginal frequency totals is equivalent to the probability value of any one cell (because there is only one degree of freedom in a 2 × 2 contingency table), determining the probability value of cell a is sufficient. If

$$P\{a+1|N-n, N-n', N\} = P\{a|N-n, N-n', N\} \times f(a)$$

then, solving for $f(a)$ produces

$$\begin{aligned} f(a) &= \frac{P\{a+1|N-n, N-n', N\}}{P\{a|N-n, N-n', N\}} \\ &= \frac{a!\, b!\, c!\, d!}{(a+1)!\, (b-1)!\, (c-1)!\, (d+1)!} \end{aligned}$$

and, after cancelling, yields

$$f(a) = \frac{(b)(c)}{(a+1)(d+1)} \ .$$

Yates provided an example analysis based on data from Milo Hellman on bottle feeding and malocclusion that had been published in *Dental Cosmos* in 1914 [609]; the data are summarized in Table 2.5 and the six exhaustive 2 × 2 contingency tables from the data in Table 2.5 are listed in Table 2.6. Yates generated the entire exact probability distribution as follows. The probability of obtaining zero normal breastfed babies for cell arrangement (1) in Table 2.6 was given by

$$P\{a=0|20, 5, 42\} = \frac{5!\ 37!\ 20!\ 22!}{42!\ 0!\ 20!\ 5!\ 17!} = 0.030957$$

Table 2.5 Hellman's data on breast feeding and malocclusion.

Feeding type	Normal teeth	Malocclusion	Total
Breast-fed baby	4	16	20
Bottle-fed baby	1	21	22
Total	5	37	42

Table 2.6 Six possible arrangements of cell frequencies with $n = 42$ and marginal frequency totals of 20, 22, 5, and 37

(1)		(2)		(3)		(4)		(5)		(6)	
0	20	1	19	2	18	3	17	4	16	5	15
5	17	4	18	3	19	2	20	1	21	0	22

and calculated utilizing a table of factorials. Then, the probability values for $a = 1, 2, 3, 4$, and 5 in Table 2.6 were recursively given by

$$P\{a = 1|20, 5, 42\} = 0.030957 \times \frac{(20)(5)}{(1)(18)} = 0.171982\ ,$$

$$P\{a = 2|20, 5, 42\} = 0.171982 \times \frac{(19)(4)}{(2)(19)} = 0.343965\ ,$$

$$P\{a = 3|20, 5, 42\} = 0.343964 \times \frac{(18)(3)}{(3)(20)} = 0.309568\ ,$$

$$P\{a = 4|20, 5, 42\} = 0.309568 \times \frac{(17)(2)}{(4)(21)} = 0.125301\ ,$$

and

$$P\{a = 5|20, 5, 42\} = 0.125301 \times \frac{(16)(1)}{(5)(22)} = 0.018226\ ,$$

respectively. In this manner, Yates was able to recursively generate the entire discrete permutation distribution from $\min(a) = \max(0, N-n-n') = \max(0, -17) = 0$ to $\max(a) = \min(N-n, N-n') = \min(20, 5) = 5$.

2.6.1 Calculation with an Arbitrary Initial Value

To illustrate the use of an arbitrary origin in a recursion procedure, consider arrangement (1) in Table 2.6 and set $C\{a = 0|20, 5, 42\}$ to some small arbitrarily-chosen value, say 5.00; thus, $C\{a = 0|20, 5, 42\} = 5.00$. Then,

$$\begin{aligned}
C\{a = 1|20,5,42\} &= 5.000000 \times \frac{(20)(5)}{(1)(18)} = 27.777778\ ,\\
C\{a = 2|20,5,42\} &= 27.777778 \times \frac{(19)(4)}{(2)(19)} = 55.555556\ ,\\
C\{a = 3|20,5,42\} &= 55.555556 \times \frac{(18)(3)}{(3)(20)} = 50.000000\ ,\\
C\{a = 4|20,5,42\} &= 50.000000 \times \frac{(17)(2)}{(4)(21)} = 20.238095\ ,
\end{aligned}$$

and

$$C\{a = 5|20,5,42\} = 20.238095 \times \frac{(16)(1)}{(5)(22)} = 2.943723\ ,$$

for a total of $C\{0,\ldots,5|20,5,42\} = 161.515152$. The desired probability values are then obtained by dividing each relative probability value by the recursively-obtained total 161.515152; e.g.,

$$\begin{aligned}
P\{a = 0|20,5,42\} &= \frac{5.000000}{161.515152} = 0.030957\ ,\\
P\{a = 1|20,5,42\} &= \frac{27.777778}{161.515152} = 0.171982\ ,\\
P\{a = 2|20,5,42\} &= \frac{55.555556}{161.515152} = 0.343965\ ,\\
P\{a = 3|20,5,42\} &= \frac{50.000000}{161.515152} = 0.309568\ ,\\
P\{a = 4|20,5,42\} &= \frac{20.238095}{161.515152} = 0.125301\ ,
\end{aligned}$$

and

$$P\{a = 5|20,5,42\} = \frac{2.943723}{161.515152} = 0.018226\ .$$

In this manner, the entire analysis could be conducted utilizing an arbitrary initial value and a recursion procedure, thereby eliminating all factorial expressions. When $\max(a) - \min(a) + 1$ is large, the computational savings can be substantial.

The historical significance of Yates' 1934 article has surely been underrated. It not only provided one the earliest and clearest explanations of Fisher's exact probability test, but also formally proposed the continuity correction to the chi-squared test for the first time. In addition, Yates' numerical studies in the paper were the first in a long and often contentious series of investigations into the best methods of testing for association in contingency tables [633, p. 17].

2.7 Irwin and Fourfold Contingency Tables

Fisher's exact probability test for 2 × 2 contingency tables was independently developed R.A. Fisher in 1935 [452], Frank Yates in 1934 [1472] and Joseph Irwin in 1935 [674]. Thus, the test is variously referred to as the Fisher exact probability test (FEPT), the Fisher–Yates exact probability test, and the Fisher–Irwin exact probability test.[22]

J.O. Irwin

It is not uncommon to find Fisher's exact probability test referred to as the Fisher–Irwin test, e.g., [33, 239, 281, 897, 1349]. Joseph Oscar Irwin earned his undergraduate degree from Christ's College, University of Cambridge, in 1921, whereupon he was offered a position with Karl Pearson at the Galton Biometric Laboratory, University College, London, with whom he had worked prior to entering Cambridge. While at University College, Irwin was in contact not only with Karl Pearson, but also with Egon Pearson and with Jerzy Neyman who was at University College, London, from 1925 to 1927 and again from 1934 to 1938. Irwin's academic degrees continued with a M.Sc. degree from the University of London in 1923, an M.A. degree from the University of Cambridge in 1924, a D.Sc. degree from the University of London in 1929 and the D.Sc. degree from the University of Cambridge in 1937 [31, 32, 550].

In 1928 Irwin joined R.A. Fisher's Statistical Laboratory at the Rothamsted Experimental Station, thereby becoming one of the few people to have studied with both Pearson and Fisher [81]. In 1931 Irwin joined the staff of the Medical Research Council at the London School of Hygiene & Tropical Medicine, where he remained for the next 30 years, except for the war years (1940–1945) when the staff of the London School of Hygiene & Tropical Medicine was evacuated from London and Irwin was temporarily attached to the Faculty of Mathematics at Queen's College, University of Cambridge, where he taught statistics to mathematicians. In his later years, Irwin was a visiting professor at the University of North Carolina at Chapel Hill during the academic years 1958–1959 and 1961–1962, and for one semester in 1965 [31]. Joseph Oscar Irwin retired in 1965 and passed away on 27 July 1982 at the age of 83 [81].

[22] Good has argued that the test should more properly be referred to as the Fisher–Yates–Irwin–Mood test [519, p. 318].

Table 2.7 Irwin's data on 2 × 2 contingency tables with equal marginal totals.

Table with 2 marked items				Table with r marked items			
Sample	Marked	Unmarked	Total	Sample	Marked	Unmarked	Total
1	2	4	6	1	r	$6-r$	6
2	6	0	6	2	$8-r$	$r-2$	6
Total	8	4	12	Total	8	4	12

In 1935 Irwin published an exact probability test for 2 × 2 contingency tables in the Italian journal *Metron* [674].[23] The publication was original and independent of the results published by Yates in 1934 [1472] and Fisher in 1935 [452] on the same theme.[24] In fact, Irwin noted in this paper that the paper was actually finished in May of 1933, but publication was "unavoidably delayed" until 1935.[25] In a footnote to this article Irwin acknowledged that a paper dealing with the same subject, "in some respects more completely" had previously been published by F. Yates in 1934.[26] In this 1935 paper Irwin described the difficulty in analyzing 2 × 2 contingency tables with Pearson's chi-squared statistic when the expected frequency in any cell was less than 5. In response to this difficulty, Irwin developed three approaches to analyze 2 × 2 contingency tables, in addition to the usual chi-squared analysis. He dismissed the first two approaches as impractical or inaccurate and advocated the third approach based on fixed marginal frequency totals [674]. An example will serve to illustrate Irwin's approach.

Consider the 2×2 contingency table on the left side of Table 2.7. Irwin observed that, given the marginal frequency totals, the cell frequency for the Marked items in Sample 1 could not be smaller than $\max(0, 6 + 8 - 12) = 2$ nor larger than $\min(6, 8) = 6$. He suggested taking samples of size 6 from a universe in which p is the probability of a Marked item. Then, the chance of getting eight Marked and four Unmarked items was

$$\binom{12}{4} p^8 (1-p)^{12-8}$$

[23] Although *Metron Rivista Internazionale di Statistica* was published in Italy, the article by Irwin was in English.

[24] For the early history of Fisher, Yates, Irwin, and the exact analysis of 2 × 2 contingency tables, see articles by Barnard [71] and Good [519–521].

[25] Irwin suffered from chronic poor health from early childhood and it is possible that was what delayed publication.

[26] Irwin joined the Rothamsted Experimental Station in 1928 and remained there until 1931, which was when Yates joined Rothamsted. Since they were both employed in Fisher's Statistical Laboratory at Rothamsted and both overlapped as undergraduates at the University of Cambridge, it is likely they were well acquainted.

and he could easily enumerate the 2 × 2 contingency tables which satisfied this condition by supposing r items in Sample 1 to be Marked, as illustrated on the right side of Table 2.7. Irwin calculated that the chance of obtaining the 2×2 table on the right side of Table 2.7 was

$$\binom{6}{r} p^r (1-p)^{6-r} \times \binom{6}{8-r} p^{8-r} (1-p)^{r-2} = \binom{6}{r}\binom{6}{8-r} p^8 (1-p)^{12-8}$$

and he then generated the probability values for all possible tables with $r = 2, \ldots, 6$; viz.,

$$\binom{6}{2}\binom{6}{6} p^8 (1-p)^{12-8} = \ \ 15 p^8 (1-p)^4 \ ,$$
$$\binom{6}{3}\binom{6}{5} p^8 (1-p)^{12-8} = 120 p^8 (1-p)^4 \ ,$$
$$\binom{6}{4}\binom{6}{4} p^8 (1-p)^{12-8} = 225 p^8 (1-p)^4 \ ,$$
$$\binom{6}{5}\binom{6}{3} p^8 (1-p)^{12-8} = 120 p^8 (1-p)^4 \ ,$$

and

$$\binom{6}{6}\binom{6}{2} p^8 (1-p)^{12-8} = \ \ 15 p^8 (1-p)^4 \ ,$$

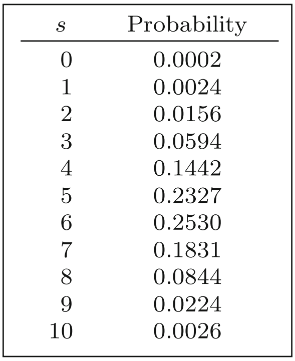
thus yielding a total of

$$\binom{12}{4} p^8 (1-p)^{12-8} = 495 p^8 (1-p)^4 \ .$$

Thus, as Irwin illustrated, if $r = 2$ the exact chance of a contingency table arising with a number of Marked items as small or smaller than in Sample 1 was $15/495 = 0.0303$ and the exact chance of an equally probable or less probable table arising was $15/495 + 15/495 = 0.0606$. Irwin then compared these results to a conventional chi-squared probability value where $\chi^2 = 6.00$, $\chi = 2.4495$, and the corresponding probability values, obtained from a $N(0, 1)$ distribution, were

Table 2.8 Irwin's data on 2 × 2 contingency tables with unequal marginal totals.

Table with 3 unmarked items				Table with s unmarked items			
Sample	Marked	Unmarked	Total	Sample	Marked	Unmarked	Total
1	79	3	82	1	$82 - s$	s	82
2	56	7	63	2	$53 + s$	$10 - s$	63
Total	135	10	145	Total	135	10	145



s	Probability
0	0.0002
1	0.0024
2	0.0156
3	0.0594
4	0.1442
5	0.2327
6	0.2530
7	0.1831
8	0.0844
9	0.0224
10	0.0026

Fig. 2.2 Probability values for the unmarked items on the right side of Table 2.8

0.0072 and 0.0143, respectively.[27] Irwin concluded that the chi-squared test would "considerably overestimate the significance" [674, p. 86] and recommended that when the numbers in all cells were small the exact method should be used, but if samples were of reasonable size and there were small cell frequencies in only one or two cells yielding expected frequencies less than five, then the researcher "shall seldom be misled by applying the usual [chi-squared] test" [674, p. 94].

Irwin concluded the article with a number of examples. In several of the examples, the row marginal frequency totals were not equal, as they are in Table 2.7 where the marginal row totals for Samples 1 and 2 are both 6. Here Irwin did something interesting and somewhat controversial, even today. A second example will illustrate that procedure.

Irwin noted that s Unmarked items in Sample 1 on the right side of Table 2.8 could take on the values 0, 1, …, 10 and he found the corresponding probability values listed in Fig. 2.2. In calculating the two-tailed probability value, Irwin noted that the observed cell frequency of 3 with a point-probability value of 0.0594 appeared in the lower tail of the distribution. He therefore accumulated all the probability values in the lower tail that were equal to or less than the observed probability value of 0.0594 to get the one-tail cumulative probability value, e.g.,

[27] To clarify, Irwin took the positive square root of χ^2, i.e., χ, which with one degree of freedom is a normal deviate, and thus obtained the probability values from a standard unit-normal table of probability values.

$0.0002 + 0.0024 + 0.0156 + 0.0594 = 0.0776$. Then Irwin calculated the upper-tail probability value as the sum of the probability values in the upper tail that were less than or equal to the observed probability value of 0.0594, e.g., $0.0224 + 0.0026 = 0.0250$. Following that, he combined the two cumulative probability values to compute $0.0776 + 0.0250 = 0.1026$ as the two-tailed probability value, whereas it was customary at the time to simply double the lower-tail probability value, i.e., $0.0776 + 0.0776 = 0.1552$. This became known as "Irwin's rule" and is still referred to today as such; see for example, Armitage and Berry [33, pp. 131–132] and Campbell [239].[28] Incidentally, Irwin's rule extends to any r-way contingency table.

2.8 The Rothamsted Manorial Estate

The Rothamsted Experimental Station began as the Rothamsted manorial estate, which can be dated from the early 1300s, when it was held by the Cressy family for about 200 years.

Manorial Estates

The manorial or seignorial system was a social and economic system of medieval Europe under which serfs and peasants tilled the arable land of a manorial estate in return for dues in kind, money, or services. A typical manorial estate was comprised of the manor house of the Lord of the Manor; the demesne, or land held and controlled by the Lord of the Manor usually consisting of arable lands, meadows, woodlands, and fish ponds; the serf holdings that were usually strips of arable land, not necessarily adjacent, which passed down through generations of serf families; and free peasants who farmed land on the estate and paid rent to the Lord of the Manor.

The meadows were usually held in common, but the woodlands and fish ponds belonged to the Lord. Serfs were expected to recompense the Lord for hunting in the woods, fishing in the ponds, and cutting wood for fuel. The Lord of the Manor collected payments from the serfs and peasants and in turn rendered protection, administered justice, and provided for the serfs in times of poor harvest [1278].

[28] The controversy as to whether to use the doubling rule or Irwin's rule to obtain a two-tailed probability value persisted for many years; see for example, articles by Cormack [279, 280] in 1984 and 1986, Cormack and Mantel in 1991 [281], Healy in 1984 [604], Jagger in 1984 [678], Mantel in 1984 and 1990 [884, 885], Yates in 1984 [1476], and Neuhäuser in 2004 [1031].

Like many other English manorial estates, Rothamsted Manor goes back to a remote antiquity [1209, p. 161].[29] Around the first century BC, the Celts occupied the Rothamsted area, leaving some archaeological evidence consisting of hearths, pot boilers, and broken pottery (i.e., shards). Under Roman rule, from about 55 BC to AD 450, Rothamsted flourished with a shrine, a flint wall around a square enclosure, and burial sites; see, for this historical period, a report by Lowther [848, p. 108–114]. The Romans left in the fifth century and were replaced by the Saxons, who left no building at the site, but gave the place its name, "Rochamstede," meaning "rook-frequented homestead" [860, 1209].

The first recorded mention of Rothamsted was in 1212 when Richard de Merston held lands there. A house with a chapel and garden are referred to in 1221 when Henry Gubion granted some land to Richard de Merston. At this time the house was a simple timber-framed building. At the beginning of the fourteenth century, Rothamsted was held by the Noels (or Nowells) who passed it to the Cressy (or Cressey) family in 1355 [542, 1352]. The Cressy family held the estate until 1525, but the male lineage died out. The Cressy's daughter, Elizabeth, remained in possession, marrying Edmund Bardolph who improved the manor house and extended the estate, purchasing the adjoining Hoos manor, among others. By the end of the sixteenth century, Rothamsted Manor was a substantial dwelling of at least 16 rooms [1352].

The Wittewronges[30] were Flemish Calvinists who, led by Jacques Wittewronge (1531–1593), emigrated from Ghent in 1564 owing to the religious persecution of Protestants by Philip II in the Spanish Netherlands at the time [574]. Jacques Wittewronges had two sons: Abraham and Jacob. Jacob Wittewronge (1558–1622) was a successful businessman and in 1611 he obtained a mortgage on Rothamsted Manor by means of a loan to Edmund Bardolph. Jacob Wittewronge married twice; his second wife was Anne Vanacker, the daughter and co-heiress of another Flemish refugee, Gerard (or Gerrard) van Acker (or Vanacker) a merchant from Antwerp who had settled in England. Anne bore Jacob Wittewronge a daughter. Anne, in 1616 and a son, John, in 1618. Jacob Wittewronge died on 22 July 1622. After Jacob's death, Anne Wittewronge married Sir Thomas Myddleton,[31] Lord Mayor of London, and in 1623 Dame Anne Myddleton procured the Rothamsted estate for her son John.

Upon the passing of Dame Anne Myddleton in 1649, John Wittewronge inherited the estate and made many improvements, especially to the manor house, holding the estate until his death on 23 June 1693. John had graduated from Trinity College, Oxford, in 1634 and by the time he was 18 had taken up his duties as Lord of the Manor [1352]. In 1640 he was knighted by Charles I. The Wittewronge descendants held the estate until male descendants ceased in 1763 and the estate then passed to

[29] For this section of the book, the authors are indebted to Sir E. John Russell (q.v. page 57) who, in 1942, compiled the early history of the Rothamsted Manor.

[30] Originally, Wittewronghele.

[31] Sometimes spelled Midleton or Middleton.

the Bennet family by the marriage of Elizabeth Wittewronge to Thomas Bennet, and finally to the Lawes family by the marriage of Mary Bennet, great-granddaughter of James Wittewronge, son of John and Elizabeth Myddleton Wittewronge, to Thomas Lawes. His son, John Bennet Lawes, was the father of John Bennet Lawes [1211, 1228, 1415]. John Bennet Lawes was born in 1814 and educated at Eton and the University of Oxford. Somehow, as a youth, he had acquired a proclivity for conducting chemical experiments, which he did at home. His early experiments were with drugs and he grew many medicinal plants on the estate, including poppies, hemlock, henbane, colchicum, and belladonna. He soon began to apply chemistry to agriculture and discovered the value of superphosphate of lime as a fertilizer and established a factory to produce the first mineral fertilizer.[32] In the 1830s Lawes established the Rothamsted Experimental Station on the estate.

Lawes died on 31 August 1900 at the age of 85 and was succeeded by his son, Charles Bennet Lawes, then aged 57, who assumed the ancestral name of Wittewronge. Unfortunately, Charles died in 1911 after a brief illness and the income had been sufficiently reduced that the family could no longer live at Rothamsted. The estate was leased to and carefully tended by Major R.B. Sidebottom and his wife, the Honorable Mrs. Sidebottom [1209, p. 166]. The Rothamsted estate was sold by the Wittewronge–Lawes family to the Rothamsted Agricultural Trust in 1934.

J.B. Lawes

John Bennet Lawes, 1st Baronet, F.R.S., Lord of Rothamsted Manor, was born on 28 December 1814 and in 1822 at the age of eight inherited his father's sixteenth century estate of somewhat more than 1,000 acres (approximately 1.7 square miles). Lawes was educated at Eton and at Brasenose College, University of Oxford, leaving in 1835 without taking a degree, whereupon he entered into the personal management of the home farm at Rothamsted of about 250 acres. In the 1830s Lawes created the Rothamsted Experimental Station on the family estate to investigate the effects on the soil of different combinations of bonemeal, burnt bones, and various types of mineral phosphate treated with sulphate or muriate of ammonia. Initially, Lawes created superphosphate from sulphuric acid and ground-up bones, then graduated to mineral phosphates, such as coprolites, and finally used imported apatite, i.e., calcium phosphate. As related by A.D. Hall, the application of sulphuric acid

(continued)

[32] Today, phosphate-based fertilizers are used throughout the world and there is presently concern that the world will eventually run out of easily accessible sources of phosphate rock [278, 784]. On the other hand, heavy spring rains generate runoff from farmer's fields into ponds and lakes, spawning growth of toxic blue-green algae, such as Microsystis aeurginosa, which are fed by the phosphorus from the fields [1463].

to calcium phosphate yields a mixture of monocalcic phosphate, phosphoric acid, and gypsum. The phosphates in this compound are soluble in water and produce an efficacious fertilizer [574, p. xxii].

On 23 May 1842 Lawes was granted a patent for the development and manufacture of superphosphate-bone meal—calcium phosphate treated with sulfuric acid—as an artificial agricultural fertilizer, and in 1843 Lawes was joined by the English chemist Sir Joseph Henry Gilbert in what began a lifelong collaboration on over 100 published articles, including papers on turnip culture, the amount of water given off by plants, the fattening qualities of different breeds of sheep, the relative advantages of malted and unmalted barley as food for stock, the valuations of unexhausted manures, nitrification, experiments on the mixed herbage of permanent meadow, climate and wheat crops, composition of rain and drainage waters, nitrogen in soils, the growth of root crops for many years in succession on the same land, the rotation of crops, and many other similar agricultural topics [331]. A full account with detailed descriptions of the major Rothamsted agricultural experiments is given is *The Book of the Rothamsted Experiments* by A.D. Hall [574]. In addition, Hall lists the publications issued from the Rothamsted Experimental Station between 1843 and 1905 [574, pp. 273–285].

A factory to manufacture superphosphate of lime was established by Lawes on 1 July 1843 at Deptford Creek, London. Lawes was elected Fellow of the Royal Society in 1854, in 1877 the University of Edinburgh conferred upon Lawes the honorary degree of LL.D., in 1882 Lawes was made a baronet, and in 1894 the University of Cambridge awarded Lawes the degree of D.Sc. Sir John Bennet Lawes F.R.S. passed away on 31 August 1900 at Rothamsted Manor at the age of 86 [331].

J.H. Gilbert

Joseph Henry Gilbert was born at Kingston-upon-Hull on 1 August 1817. He was educated at Glasgow University where he worked in the laboratory of Professor Thomas Thomson. He moved to University College, London, in the autumn of 1839 and worked briefly in the laboratory of Professor Anthony Todd Thomson. It was in Thomson's laboratory that Gilbert and Lawes first met. He received his Ph.D. in 1840 from the University of Giessen in Germany where he studied under the renowned chemist, Professor Justus van Liebig, who had established the world's first major school of chemistry. Another famous student of von Liebig was August Kekulé, the discover of the benzene ring [1180, pp. 133–135].

(continued)

Gilbert, at the age of 26, was invited by Lawes on 1 June 1843 to oversee the Rothamsted experiments. Thus began a partnership in research that lasted for 58 years. Lawes possessed an originating mind and had a thorough knowledge of practical agriculture. Gilbert, on the other hand, was possessed of indomitable perseverance, combined with extreme patience. In his research he united scrupulous accuracy with attention to detail. In general, Lawes directed the agricultural operations in the experimental fields and the execution of the experiments was in the hands of Gilbert [574, pp. xxii–xl]. Gilbert was elected Fellow of the Royal Society in 1860 and knighted by Queen Victoria in 1893. Sir Joseph Henry Gilbert F.R.S. died at his home in Harpenden on 23 December 1901 in his 85th year and is buried in the churchyard of St. Nicholas Church, next to his long-time friend, John Bennet Lawes [184, 1416].

The Experimental Station

The Rothamsted Experimental Station, now Rothamsted Research, in Harpenden, Hertfordshire, England, about 25 miles northeast of London, had its beginnings in the 1830s, *vide supra*. Together Lawes and Gilbert established the Rothamsted Experimental Station on the family estate, the first agricultural research station in the world, and in 1889 Lawes established the Lawes Agricultural Trust, setting aside £100,000, one-third of the proceeds from the sale of his fertilizer business in 1872, to ensure the continued existence of the Rothamsted Experimental Station [184,331,1280] (According to the Rothamsted Research website, the equivalent amount today would be approximately £5,000,000 or $7,800,000 [341].) In 1911 David Lloyd George, Chancellor of the Exchequer set up the Development Fund for the rehabilitation of British farming, making £1,000,000 available for research funding. In 1867 Lawes and Gilbert received the Royal Society's Royal Medal, also called the Queen's medal, awarded for important contributions in the applied biological and physical sciences.

Expansions beginning in 1902 provided new facilities and added chemists, bacteriologists, and botanists to the staff at Rothamsted. Researchers at Rothamsted have made many significant contributions to science over the years, including the discovery and development of the pyrethroid insecticides, as well as pioneering contributions in the fields of virology, nematology, soil science, and pesticide resistance. In 2012 Rothamsted Research supported 350 scientists, 150 administrative staff, and 60 Ph.D. students [341].

Sir John Russell, who came from Wye Agricultural College[33] in 1907 and assumed the directorship of the Rothamsted Experimental Station in 1912, appointed R.A. Fisher to the Rothamsted Experimental Station in October, 1919 and commissioned him to study yield data on 67 years of Broadbalk wheat,[34] for which trials had begun as far back as 1843. Sir Russell initially hired Fisher on a temporary basis, as he had only £200 appropriated for the appointment, but he soon recognized the genius of Fisher and set about securing the necessary funds to hire him on a permanent basis; however, not before Fisher had spent twice the £200 [191, p. 792]. Fisher made Rothamsted into a major center for research in statistics and genetics, remaining at Rothamsted as the head of the Statistical Laboratory until 1933 when he left to assume the post of Galton Professor of Eugenics at University College, London. Fisher was succeeded by Frank Yates who had come to Rothamsted in 1931 as Assistant Statistician. Regular afternoon tea had been instituted at Rothamsted in 1906, 13 years prior to Fisher's arrival, when Dr. Winifred E. Brenchley joined the scientific staff as its first woman member [1354].[35] Sir John Russell recalled:

> [n]o one in those days knew what to do with a woman worker in a laboratory; it was felt, however, that she must have tea, and so from the day of her arrival a tray of tea and a tin of Bath Oliver biscuits appeared each afternoon at four o'clock precisely; and the scientific staff, then numbering five, was invited to partake thereof [1210, p. 235] (Russell, quoted in Box [195, p. 132]).

This tea service ended up being an important part of the story of Fisher and the beginnings of permutation methods.

E.J. Russell

Edward John Russell was born on 31 October 1872 and was educated at Carmarthen Presbyterian College, Aberystwyth University College, and Owen's College, Manchester, graduating with a B.Sc. and First Class Honors in Chemistry in 1896. Russell was awarded the degree of D.Sc. by the University of London for his researches at Manchester [1195, 1361].

In January 1901 Russell, who preferred the name John Russell, obtained a Lectureship in Chemistry at Wye Agricultural College, at which the Principal

(continued)

[33] The College of St. Gregory and St. Martin at Wye, more commonly known as Wye College, was an educational institution in the small village of Wye, Kent, about 60 miles east of London.

[34] Broadbalk refers to the fields at Rothamsted on which winter wheat was cultivated, not a strain of wheat.

[35] Afternoon tea had been a British tradition since one of Queen Victoria's (1819–1901) ladies-in-waiting, Anna Maria Russell (née Stanhope) (1783–1857), the seventh Duchess of Bedford, introduced it at Belvoir (pronounced Beaver) Castle in the summer of 1840, the idea being a light repast around 4 p.m. would bridge the lengthy gap between luncheon and dinner, which in fashionable circles at that time was not taken until 8 p.m.

was Alfred Daniel Hall. Hall left Wye shortly after Russell joined the staff to become Director of Rothamsted Experimental Station. Meanwhile, the Goldsmith's Company had given a capital grant of £10,000 to endow a position in soil research at Rothamsted, which allowed Hall and the Lawes Agricultural Trust to offer Russell a post as the first Goldsmith's Company Soil Chemist. Russell accepted the offer and moved from Wye College to Rothamsted in July of 1907. At that time the scientific staff was comprised of Hall and Russell and, in addition, Winifred Elsie Brenchley as botanist, Henry Brougham Hutchinson as bacteriologist, and Norman H.J. Miller as chemist [1361, 1404].

Hall left Rothamsted in October of 1912 and Russell was appointed Director of the Rothamsted Experimental Station in 1912 and served as Director until 1943. He was elected Fellow of the Royal Society in 1917, received the Order of the British Empire in 1918, and was knighted by King George V in 1922. In 1943, Russell, now 70, retired from Rothamsted and was succeeded by William Gammie Ogg. Sir E. John Russell O.B.E. F.R.S. died on 12 July 1965 at the age of 92. A complete bibliography of his writings and publications is contained in a biography by Thornton [1361, pp. 474–477].

In *The Design of Experiments* (familiarly known as *DOE*), first published in 1935, Fisher (q.v. page 25) again intimated at the utility of a permutation approach to obtain exact probability values [451, Sect. 11], and it is this formative text that many researchers refer to as setting the idea of permutation tests into motion, e.g., Conover [272], Kempthorne [719], Kruskal and Wallis [779], and Wald and Wolfowitz [1407]. Fisher's description of the "lady tasting tea" is often referenced to describe the underlying logic of permutation tests. It appears that the story has never been told in its entirety in a single place and is worth relating. While several versions of the story exist, the account here relies primarily on the description by Joan Fisher Box [195, pp. 131–132].

2.8.1 The Rothamsted Lady Tasting Tea Experiment

The "lady tasting tea" experiment at the Rothamsted Experimental Station in the early 1920s has become one of the most referenced experiments in the statistical literature. A search of the Internet in February of 2013 produced 25,600 citations.[36]

[36] For a detailed explanation as to why it matters whether the tea or the milk is poured into the teacup first, see a 2012 article by Stephen Senn in *Significance* [1251].

The Lady Tasting Tea

At Rothamsted in the 1920s, afternoon tea was served at 4 o'clock in the sample house in inclement weather or, otherwise, outside the sample house on a table set with an urn of tea and cups. One afternoon in the early 1920s, Fisher drew a cup of tea from the urn, added milk, and proffered it to the lady beside him, Dr. Blanche Muriel Bristol, an algologist. She declined the cup of tea offered by Fisher, stating that she preferred a cup into which the milk had been poured first. Fisher's quick response was, "[n]onsense, surely it makes no difference" [195, p. 134].

Dr. William A. Roach, a chemist at the laboratory who was soon to marry Dr. Bristol, suggested a test, to which Dr. Bristol agreed. Consequently, eight cups of tea were prepared, four with the tea added after the milk and four with the milk added after the tea, and presented to Dr. Bristol in random order [195, p. 134]. Dr. Bristol's personal triumph was never recorded and Fisher does not describe the outcome of the experiment; however, H. Fairfield Smith was present at the experiment and he later reported that Dr. Bristol had identified all eight cups of tea correctly [1218, p. 8]. William Roach, however, apparently reported that Dr. Bristol "made nearly every choice correctly" [191, p. 793]. Incidentally, the probability of correctly dividing the eight cups into two sets of four by chance alone is only 1 in 70 or 0.0143. It should be noted that another version of the story has the event taking place at the University of Cambridge in the late 1920s [1218], but it seems unlikely that this version of the story is correct. In addition, according to Dr. Roach, Dr. Bristol was correct on enough of the cups to prove her point [575, 1251].[37]

For additional descriptions of the tea tasting experiment, see Fisher [451, pp. 11–29], Fisher [459, Chap. 6], Box [191], Box [195, pp. 134–135], Gridgeman [555], Salsburg [1218, pp. 1–2], Lehmann [816, pp. 63–64], Hall [575, p. 315], Okamoto [1053], Senn [1250–1252], and Springate [1313]. For a decidedly different (Baysian) take on the lady tasting tea experiment, see a 1984 paper on "A Bayesean lady tasting tea" by Dennis Lindley [829] and a 1992 paper on "Further comments concerning the lady tasting tea or beer: P-values and restricted randomization" by Irving (I.J.) Good [521].

[37] For a biography of Dr. B. Muriel Bristol and a picture, see a 2012 article by Stephen Senn in *Significance* [1251].

Table 2.9 Five possible arrangements of cell frequencies with $n = 8$ and identical marginal frequency totals of 4, 4, 4, and 4

(1)		(2)		(3)		(4)		(5)	
4	0	3	1	2	2	1	3	0	4
0	4	1	3	2	2	3	1	4	0

2.8.2 Analysis of The Lady Tasting Tea Experiment

A dozen years later, in 1935, Fisher provided a detailed discussion of the tea tasting experiment [451].[38] In what Fisher termed a hypothetical experiment in Chap. II, Sect. 5 of *The Design of Experiments*, Fisher described a woman who claimed to be able to tell the difference between tea with milk added first and tea with milk added second [451]. He concocted an experiment, without mentioning the Rothamsted experiment or Dr. Bristol, whereby a woman sampled eight cups of tea, four of each type, and identified the point at which the milk had been added—before the tea, or after.[39] Fisher then outlined the chances of the woman being correct merely by guessing, based on the number of trials; in this case, eight cups of tea [646]. The five possible 2 × 2 tables are listed in Table 2.9.

The null hypothesis in this experiment was that the judgments of the lady were in no way influenced by the order in which the ingredients were added. Fisher explained that the probability of correctly classifying all eight cups of tea was one in 70, i.e., the hypergeometric point-probability value for cell arrangement (1) in Table 2.9 is given by

$$P\{4|4,4,8\} = \frac{4!\;4!\;4!\;4!}{8!\;4!\;0!\;4!\;0!} = \frac{24}{1{,}680} = \frac{1}{70}\ .$$

Fisher went on to note that only if every cup was correctly classified would the lady be judged successful; a single mistake would reduce her performance below the level of significance. For example, with one misclassification the one-tailed probability for cell arrangements (1) and (2) in Table 2.9 is given by

$$P\{3|4,4,8\} + P\{4|4,4,8\} = \frac{4!\;4!\;4!\;4!}{8!\;3!\;1!\;3!\;1!} + \frac{4!\;4!\;4!\;4!}{8!\;4!\;0!\;4!\;0!} = \frac{16}{70} + \frac{1}{70} = \frac{17}{70}$$

and $17/70 = 0.2429$ is much greater than 0.05, whereas $1/70 = 0.0143$ is considerably less than 0.05.

[38] In 1956 Fisher published a lengthy discussion of the lady tasting tea experiment titled "Mathematics of a lady tasting tea" in J.R. Newman's book titled *The World of Mathematics* [459, pp. 1512–1521].

[39] It should be noted that Francis Galton, after much experimentation, always chose to put the milk into the teacup first [1251, p. 32].

Table 2.10 Seven possible arrangements of cell frequencies with $n = 36$ and identical marginal frequency totals of 6, 6, 6 and 6

(1)		(2)		(3)		(4)		(5)		(6)		(7)	
0	6	1	5	2	4	3	3	4	2	5	1	6	0
6	0	5	1	4	3	3	3	2	4	1	5	0	6

To increase the sensitivity of the experiment, Fisher suggested a new experiment with 12 cups of tea, six with the milk added first and six with the milk added second. Table 2.10 lists the seven possible 2 × 2 tables. Here the hypergeometric probability of correctly classifying all 12 cups of tea as listed in cell arrangement (1) of Table 2.10 is one in 924 and is given by

$$P\{0|6,6,12\} = \frac{6!\;6!\;6!\;6!}{12!\;0!\;6!\;6!\;0!} = \frac{720}{665{,}280} = \frac{1}{924}\;,$$

and for one misclassification the one-tailed probability for cell arrangements (1) and (2) in Table 2.10 is given by

$$\begin{aligned} &P\{1|6,6,12\} + P\{0|6,6,12\} \\ &\qquad = \frac{6!\;6!\;6!\;6!}{12!\;1!\;5!\;5!\;1!} + \frac{6!\;6!\;6!\;6!}{12!\;0!\;6!\;6!\;0!} = \frac{36}{924} + \frac{1}{924} = \frac{37}{924}\;. \end{aligned}$$

Fisher determined that since $37/924 = 0.04$ was less than 0.05, the experiment would be considered significant even with one misclassification. This additional configuration led Fisher to observe that increasing the size of the experiment rendered it more sensitive and he concluded that the value of an experiment is increased whenever it permits the null hypothesis to be more readily disproved. It should be noted that in this example Fisher simply assumed 0.05 as the level of significance, without explicitly identifying the level of significance.[40]

2.9 Fisher and the Analysis of Darwin's *Zea mays* Data

In 1935 Fisher (q.v. page 25) provided a second hypothetical discussion of permutation tests in *The Design of Experiments*, describing a way to compare the means of randomized pairs of observations by permutation [451, Sect. 21].

[40] It is generally understood that the conventional use of the 5 % level of significance as the maximum acceptable probability for determining statistical significance was established by Fisher when he developed his procedures for the analysis of variance in 1925 [292]. Fisher also recommended 0.05 as a level of significance in relation to chi-squared in the first edition of *Statistical Methods for Research Workers* [448, pp. 79–80]. Today, $p = 0.05$ is regarded as sacred by many researchers [1281]. However, Fisher readily acknowledged that other levels of significance could be used [449, p. 504]. In this regard, see discussions by Cowles and Davis [292] and Lehmann [816, pp. 51–53].

Table 2.11 Heights of crossed- and self-fertilized *Zea mays* plants in inches

	Crossed-	Self-	Difference	Difference
Pot	fertilized	fertilized	(inches)	(eighths)
I	$23\frac{4}{8}$	$17\frac{3}{8}$	$+6\frac{1}{8}$	$+49$
	12	$20\frac{3}{8}$	$-8\frac{3}{8}$	-67
	21	20	$+1$	$+8$
II	22	20	$+2$	$+16$
	$19\frac{1}{8}$	$18\frac{3}{8}$	$+0\frac{6}{8}$	$+6$
	$21\frac{4}{8}$	$18\frac{5}{8}$	$+2\frac{7}{8}$	$+23$
III	$22\frac{1}{8}$	$18\frac{5}{8}$	$+3\frac{4}{8}$	$+28$
	$20\frac{3}{8}$	$15\frac{2}{8}$	$+5\frac{1}{8}$	$+41$
	$18\frac{2}{8}$	$16\frac{4}{8}$	$+1\frac{6}{8}$	$+14$
	$21\frac{5}{8}$	18	$+3\frac{5}{8}$	$+29$
	$23\frac{2}{8}$	$16\frac{2}{8}$	$+7$	$+56$
IV	21	18	$+3$	$+24$
	$22\frac{1}{8}$	$12\frac{6}{8}$	$+9\frac{3}{8}$	$+75$
	23	$15\frac{4}{8}$	$+7\frac{4}{8}$	$+60$
	12	18	-6	-48
Total	$302\frac{7}{8}$	$263\frac{5}{8}$	$+39\frac{2}{8}$	$+314$

In this case Fisher carried the example through for the first time, calculating test statistics for all possible pairs of the observed data [646]. For this example analysis, Fisher considered data from Charles Darwin on 15 pairs of planters containing *Zea mays* ("maize" in the United States) seeds in similar soils and locations, with heights to be measured when the plants reached a given age [318]. As Darwin described the experiment, *Zea mays* is monoecious and was selected for trial on this account.[41] Some of the plants were raised in a greenhouse and crossed with pollen taken from a separate plant; and other plants, grown separately in another part of the greenhouse, were allowed to fertilize spontaneously. The seeds obtained were placed in damp sand and allowed to germinate. As they developed, plant pairs of equal age were planted on opposite sides of four very large pots, which were kept in the greenhouse. The plants were measured to the tips of their leaves when between 1 and 2 ft in height. The data from the experiment are given in the first two columns of Table 2.11 and are from Table XCVII in Darwin's *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom* [318, p. 234].

Using the data in the last column of Table 2.11 where the differences between the heights of the crossed- and self-fertilized plants were recorded in eighths of an inch,

[41] For a concise summary of the *Zea mays* experiment, see a discussion by Erich Lehmann in his posthumously published 2011 book on *Fisher, Neyman, and the Creation of Classical Statistics* [816, pp. 65–66].

Fisher first calculated a matched-pairs t test. He found the mean difference between the crossed- and self-fertilized *Zea mays* plants to be

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{314}{15} = 20.933$$

and the standard error to be

$$s_{\bar{d}} = \sqrt{\frac{\displaystyle\sum_{i=1}^{n} d_i^2 - \bar{d}\sum_{i=1}^{n} d_i}{n(n-1)}} = \sqrt{\frac{26{,}518 - (20.933)(314)}{15(15-1)}} = 9.746 \; .$$

Then, Student's matched-pairs t test yielded an observed statistic of

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{20.933}{9.746} = 2.148 \; .$$

Fisher pointed out that the 5 % t value with 14 degrees of freedom was 2.145 and concluded since 2.148 just exceeded 2.145, the result was "significant" at the 5 % level.

Fisher then turned his attention to an exact permutation test, calculating sums of the differences for the $2^{15} = 32{,}768$ possible arrangements of the data, based on the null hypothesis of no difference between self-fertilized and cross-fertilized *Zea mays* plants. The exact probability value was calculated as the proportion of values with differences as, or more extreme, than the observed value. Fisher found that in 835 out of 32,768 cases the deviations were greater than the observed value of 314; in an equal number of cases, less than 314; and in 28 cases, exactly equal to 314. Fisher explained that in just $835 + 28 = 863$ out of a possible 32,768 cases, the total deviation would have a positive value as great or greater than the observed value of 314, and in an equal number of cases it would have as great a negative value. The two groups together constituted $1{,}726/32{,}768 = 5.267\,\%$ of the possibilities available, a result very nearly equivalent to that obtained using Student's t test, where the two-tailed probability value for $t = 2.148$ with 14 degrees of freedom is 4.970 % [461, p. 47]. Fisher additionally noted that the example served to demonstrate that an "independent check" existed for the "more expeditious methods" that were typically in use, such as Student's t test [451, pp. 45–46].

Finally, Fisher argued that, because the t distribution is continuous and the permutation distribution is discrete, the t distribution was counting only half of the 28 cases that corresponded exactly with the observed total of 314. He went on to show that making an adjustment corresponding to a correction for continuity provided a t probability value more in line with the exact probability value. The corrected value of t was 2.139, yielding a probability value of 5.054 % which is closer to the exact value of 5.267 % than the unadjusted value of 4.970 %. For excellent synopses of the *Zea mays* experiment, see discussions by Kempthorne [719, p. 947], Holschuh [646], Lehmann [816, pp. 65–66], McHugh [914], and E.S. Pearson [1093].

One of the benefits Fisher attributed to permutation methods was its utility in validating normal-theory analyses [451, Chaps. 20 and 21]. Here Fisher argued that, when testing the hypothesis of no treatment effect in an agricultural experiment, the normal-theory significance level usually approximates the corresponding permutation significance level. As noted by Hooper [647], this tendency for agreement between normal-theory and permutation tests has also been examined using both real and simulated data by Eden and Yates [379] and Kempthorne and Doerfler [725]; moment calculations by Bailey [49], Pitman [1131], and Welch [1428]; Edgeworth expansions by Davis and Speed [329]; and limit theorems by Ho and Chen [634], Hoeffding [636], and Robinson [1178]. In this regard, Fisher was fond of referring to a 1931 article by Olof Tedin [1343] in which Tedin demonstrated that when the assumptions of the classical analysis of variance test are met in practice, the classical test and the corresponding randomization test yielded essentially the same probability values [1126].

O. Tedin

Olof Tedin (1898–1966) was a Swedish geneticist who spent most of his professional career as a plant breeder with the Swedish Seed Association, Svalöf, where he was in charge of the breeding of barley and fodder roots in the Weibullsholm Plant Breeding Station, Landskrona. In 1931, with the help of Fisher, he published a paper on the influence of systematic plot arrangements on the estimate of error in field experiments [1343]. Fisher had previously shown that of the numerous possible arrangements of plots subject to the condition that each treatment should appear once in each row and once in each column (an Euler Latin Square), it was possible to choose at random one to be used in the field that would be statistically valid. Tedin fashioned 12 blocks of 5 × 5 plots with five treatments distributed according to different plans.

Two of the 12 arrangements were knight's moves (Knut Vik), Latin Squares in which all cells containing any one of the treatment values can be visited by a succession of knight's moves (as in chess) and where no two diagonally adjacent cells have the same treatment value; two of the arrangements were diagonal Latin Squares in which each of the treatment values appears once in one of the diagonals and the other diagonal is composed of the same treatment value, e.g., all 1s; seven of the arrangements were random arrangement Latin Squares, as recommended by Fisher [449]; and one was a specially constructed Latin Square to evaluate "spread," wherein arrangements in which adjacent plots never have the same treatment.

(continued)

Examples of the knight's move, diagonal, and random Latin Square arrangements used by Tedin are:

3	4	5	1	2
5	1	2	3	4
2	3	4	5	1
4	5	1	2	3
1	2	3	4	5

Knight's Move

2	3	4	5	1
3	4	5	1	2
4	5	1	2	3
5	1	2	3	4
1	2	3	4	5

Diagonal

4	3	1	5	2
1	5	2	3	4
5	2	4	1	3
2	1	3	4	5
3	4	5	2	1

Random

Tedin found that systematic arrangements introduced bias in the estimate of the error of the experiment, with the knight's move arrangements over-estimating the error and the diagonal arrangements under-estimating the error. He concluded that "the present study confirms the views of Fisher, not only in the one special case, but in all other cases of systematic plot arrangements as well" [1343, p. 207].

2.10 Fisher and the Coefficient of Racial Likeness

Fisher's 1936 article on "'The coefficient of racial likeness' and the future of craniometry" provided an alternative explanation of how permutation tests work [453]. Without explicitly labeling the technique a permutation test, Fisher described a shuffling procedure for analyzing data. His description began with two hypothetical groups of $n_1 = 100$ Frenchmen and $n_2 = 100$ Englishmen with a measurement of stature on each member of the two groups. After recording the differences in height between the two groups in the observed data, the measurements were recorded on 200 cards, shuffled, and divided at random into two groups of 100 each, a division that could be repeated in an enormous, but finite and conceptually calculable number of ways. [42] A consideration of all possible arrangements of the pairs of cards would provide an answer to the question, "Could these samples have been drawn at random from the same population?" [453, p. 486]. Fisher explained that a statistician usually does not carry out this tedious process, but explained that the statistician's conclusions "have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method" [453, p. 58]. Fisher went on to stress that the test of significance calculates a probability value and does not

[42] Authors' note: actually, 90,548,514,656,103,281,165,404,177,077,484,163,874,504,589,675,413, 336,841,320 ways.

calculate a metrical difference [453, pp. 59–60], anticipating perhaps the current emphasis on calculating effect sizes as well as tests of significance.

Finally, it should be noted that while Fisher never referenced nor provided a footnote to Karl Pearson in this article, it is abundantly evident that this article is a thinly-veiled criticism of Pearson's coefficient of racial likeness published in 1926 [1110], as the formula for the coefficient of racial likeness on page 60 of Fisher's article is taken directly from Pearson's 1926 article. For a concise description of the card shuffling experiment and a critical retort to Fisher's analyses of Darwin's *Zea mays* data and the racial craniometry data see E.S. Pearson [1093], a summary of which is provided on page 76.

Continuing the theme of shuffling cards to obtain permutations of observed data sets, in 1938 Fisher and Yates described in considerable detail an algorithm for generating a random permutation of a finite set, i.e., shuffling the entire set [463, p. 20]. The basic method proposed by Fisher and Yates consisted of four steps and resulted in a random permutation of the original numbers [463, p. 20]:

1. Write down all the numbers from 1 to n, where n is the size of the finite set.
2. Pick a number k between 1 and n and cross out that number.
3. Pick a number k between 1 and $n - 1$, then counting from the low end, cross out the kth number not yet crossed out.
4. Repeat step 3, reducing n by one each time.[43]

2.11 Hotelling–Pabst and Simple Bivariate Correlation

While at Columbia University, Harold Hotelling was a charter member of the Statistical Research Group (q.v. page 69) along with Jacob Wolfowitz and W. Allen Wallis. This elite membership brought him into contact with a number of talented and influential statisticians of the day.

H. Hotelling

Harold Hotelling entered the University of Washington in Seattle in 1913 but his education was interrupted when he was called up for military service in World War I. Hotelling recalled that he, "having studied mathematics, science and classics at school and college, was considered by [the] Army authorities competent to care for mules. The result was [that] a temperamental mule named Dynamite temporarily broke my leg and thereby saved his life, as

(continued)

[43] The Fisher–Yates shuffle, with little change, became the basis for more sophisticated computer shuffling techniques by Richard Durstenfeld in 1964 [367], Donald Knuth in 1969 [762], and Sandra Sattolo in 1986 [1222]. N. John Castellan [245] and Timothy J. Rolfe [1188] urged caution in choosing a shuffling routine as many widely-used shuffling algorithms are incorrect.

the rest of the division was sent to France and [was] wiped out" (Hotelling, quoted in Darnell [317, p. 57]). Hotelling was discharged from the Army on 4 February 1919, and returned to the University of Washington to continue his studies.

Hotelling earned his B.A. degree in journalism from the University of Washington in 1919, his M.S. degree in mathematics from the University of Washington in 1921, and his Ph.D. in mathematics (topology) from Princeton University under Oswald Veblen in 1924. The topic of the thesis was "Three-dimensional Manifolds of States of Motion." He began his career at Stanford University, first as a research associate with the Food Research Institute from 1924 to 1927, and then as an Associate Professor in the Department of Mathematics from 1927 to 1931. It was during this time that Hotelling began corresponding with Fisher in England. This correspondence eventually led to Hotelling traveling to the Rothamsted Experimental Station to study with Fisher in 1929. In his unsolicited review of Fisher's *Statistical Methods for Research Workers*, first published in 1925, Hotelling wrote:

> [m]ost books on statistics consist of pedagogic rehashes of identical material. This comfortably orthodox subject matter is absent from the volume under review, which summarizes for the mathematical reader the author's independent codification of statistical theory and some of his brilliant contributions to the subject, not all of which have previously been published [651, p. 412].

Despite the fact that the book did not receive even one other single positive review [576, p. 219], Hotelling concluded that Fisher's "work is of revolutionary importance and should be far better known in this country" [651, p. 412]. Hotelling was so impressed with *Statistical Methods for Research Workers* that he volunteered a review for the second edition in 1928. Hotelling subsequently volunteered a review for the third, fourth, fifth, sixth, and seventh editions [816, p. 22]. Eventually, 14 editions of *Statistical Methods for Research Workers* were published, the last in 1970, and it has been translated into six languages [192, p. 153].

Hotelling was recruited to Columbia University in 1931 as Professor of Economics and to initiate a Mathematical Statistics program. Columbia long had a reputation for incorporating statistical methods into the social sciences, especially economics under the leadership of Henry Ludwell Moore, but also in psychology with James McKeen Cattell, anthropology with Franz Boas, and sociology with Franklin Henry Giddings [238]. While at Columbia, Hotelling was a charter member of the Statistical Research Group (q.v. page 69). In 1946 Hotelling left Columbia University for the University of North Carolina at Chapel Hill at the urging of Gertrude Mary Cox to establish what would become a renowned Department of Mathematical Statistics. Harold Hotelling retired in 1966 and died on 26 December 1973 at the age of 78

(continued)

from injuries sustained after falling on a patch of ice outside his home at Chapel Hill, North Carolina [37, 814, 1058, 1288].

M.R. Pabst

Margaret Hayes Pabst (née Richards) graduated with an A.B. degree from Vassar College in 1931 [1076, p. 3], received her A.M. degree from the University of Chicago in mathematics in 1932, and earned her Ph.D. in economics from Columbia University in 1944, where she studied with Hotelling.[44] In 1935 Margaret Hayes Richards married William Richard Pabst, Jr., who was at that time teaching economics at Cornell University [826, p. 752]. In that same year, Margaret Pabst was hired as an assistant in the College of Agriculture at Cornell University [826, p. 752]. In the fall of 1936 William Pabst returned to his alma mater, Amherst College, as an Assistant Professor, and from 1936 to 1938 Margaret Pabst was employed as a researcher with the Council of Industrial Studies at Smith College in nearby Northampton, Massachusetts. Her major work for the Council was a report titled "Agricultural Trends in the Connecticut Valley Region of Massachusetts, 1800–1900," which was her dissertation at Columbia University and was later published in *Smith Studies in History* [1079]. Margaret Pabst also published a small volume in 1932 on *Properties of Bilinear Transformations in Unimodular Form* that was the title of her Master's thesis at the University of Chicago [1077], and another small volume in 1933 on *The Public Welfare Administration of Dutchess County, New York* that was the Norris Fellowship Report of 1932–1933 [1078].

In 1938 William Pabst accepted a position as Associate Professor of Economics at Tulane University in New Orleans, Louisiana [1080, p. 876] and in 1941 William and Margaret Pabst moved to Washington, DC, where he worked for the War Production Board and the Office of Price Administration until 1944, when he went into the Navy and was stationed at the Bureau of

(continued)

[44] Authors' note: special thanks to Nanci A. Young, College Archivist, William A. Neilson Library at Smith College, Northampton, Massachusetts, for retrieving the information on Margaret Richards Pabst, and to Nancy Lyons, Program Analyst, United States Department of Agriculture, Food and Nutrition Service, for contacting Archivist Nanci Young at Smith College on our behalf. Special thanks also to Sarah Jane Pabst Hogenauer and Dr. Margaret Pabst Battin, Distinguished Professor of Philosophy and Adjunct Professor of Internal Medicine, Division of Medical Ethics at the University of Utah, who are the daughters of Margaret Richards Pabst and who graciously shared details of their mother's life, including having Muriel Hotelling, Harold Hotelling's daughter, as a babysitter and, as girls of 10 or 11, having lunch with R.A. Fisher.

Ordnance in Washington, DC [1284, p. C4]. In 1946 he left active duty and became Chief Statistician in the Navy's Bureau of Ordnance as a civilian. Margaret Pabst also worked for the United States government during the war, and after the war, taught piano and published two books on music, co-authored with Laura Pendleton MacCartney. Margaret Hayes Richards Pabst died on 15 April 1962 in Washington, DC.

While at Columbia University, on 1 July 1942, Harold Hotelling along with W. Allen Wallis and Jacob Wolfowitz, became charter members of the renowned Statistical Research Group which was based at Columbia during World War II and remained in existence until 30 September 1945. The SRG attracted an extraordinary group of research statisticians to Columbia and brought Hotelling into contact with many of the foremost mathematical statisticians of the time [1219].

The SRG at Columbia

The Statistical Research Group (SRG) was based at Columbia University during the Second World War from 1942 to 1945 and was supported by the Applied Mathematics Panel of the National Defense Research Committee, which was part of the Office of Scientific Research and Development (OSRD). In addition to Harold Hotelling, Wilson Allen Wallis, and Jacob Wolfowitz, the membership of the SRG included Edward Paulson, Julian Bigelow, Milton Friedman, Abraham Wald, Albert Bowker, Harold Freeman, Rollin Bennett, Leonard Jimmie Savage, Kenneth Arnold, Millard Hastay, Abraham Meyer Girshick, Frederick Mosteller, Churchill Eisenhart, Herbert Solomon, and George Stigler [1412]. For concise histories of the SRG, see articles by W. Allen Wallis [1412] and Ingram Olkin [1056, pp. 123–125].

In 1936 Hotelling and Pabst used permutation methods for calculating exact probability values for small samples of rank data in their research on simple bivariate correlation [653]. Noting that tests of significance are primarily based on the assumption of a normal distribution in a hypothetical population from which the observations are assumed to be a random sample, Hotelling and Pabst set out to develop methods of statistical inference without assuming any particular distribution of the variates in the population from which the sample had been drawn. Hotelling and Pabst noted that a false assumption of normality usually does not give rise to serious error in the interpretation of simple means due to the central limit theorem, but cautioned that the sampling distribution of second-order statistics are more seriously disturbed by the lack of normality and pointed to "the grave dangers in using even those distributions which for normal populations are accurate, in the absence of definite evidence of normality" [653, p. 30]. Hotelling and Pabst also cautioned researchers about the pitfalls of using Pearson's standard error to provide probability values, noting that in order to use the standard error it was necessary to assume that (1) the underlying population must be distributed as bivariate normal—a more stringent assumption than requiring that each variate be normally distributed, (2) only the first few terms of Pearson's infinite series are sufficient,[45] (3) the distribution of Spearman's rank-order correlation coefficient is normal, and (4) sample values can be substituted for population values in the formula for the standard error.

Consider n individuals arranged in two orders with respect to two different attributes. If X_i denotes the rank of the ith individual with respect to one attribute and Y_i the rank with respect to the other attribute so that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are two permutations of the n natural integers $1, \ldots, n$, then define $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$ where $\bar{X} = \bar{Y} = (n+1)/2$.[46] The rank-order correlation coefficient is then defined as

$$r' = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i}{\sqrt{\displaystyle\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \ . \tag{2.2}$$

Hotelling and Pabst showed that

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} X_i\right)^2}{n}$$
$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n^3 - n}{12} \ ,$$

and $\sum_{i=1}^{n} y_i^2$ have the same value. Denote by d_i the difference between the two ranks for the ith individual, so that $d_i = X_i - Y_i = x_i - y_i$, then

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 = \frac{n^3 - n}{6} - 2\sum_{i=1}^{n} x_i y_i \ .$$

Substituting into Eq. (2.2) and simplifying yields

[45] In 1907, Pearson derived the standard error of Spearman's rank-order correlation coefficient. Assuming normality, Pearson generated the first four terms of an infinite series to provide an approximate standard error [1109].

[46] In the early years of statistics it was common to denote raw scores with upper-case letters, e.g., X and Y, and deviations from the mean scores with lower-case letters, e.g., x and y.

$$r' = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} ,$$

which is Spearman's rank-order correlation coefficient, first published by Charles Spearman in 1904 in *American Journal of Psychology* [1300].

The article by Hotelling and Pabst utilized the calculation of a probability value that incorporated all $n!$ permutations of the data, under the null hypothesis that all permutations were equally-likely (q.v. page 4).[47] The probability for any particular value was calculated as the proportion of the number of permutations equal to or more extreme than the value obtained from the observed data. Following on the work of Charles Spearman and Karl Pearson who had provided rough standard deviations for a measure of rank-order correlation, Hotelling and Pabst provided a thorough and accurate analysis that allowed for small samples. Although Hotelling and Pabst did not produce tables for tests of significance, they did provide exact probability values for small samples of $n = 2$, 3, and 4 [653, p. 35]. Finally, reflecting the frustration of many statisticians in the 1930s, Hotelling and Pabst observed that for large samples the calculation of exact probability values was very laborious, forcing researchers to use approximations.

It is notable that while earlier works contained the essence of permutation tests, the article by Hotelling and Pabst included a much more explicit description of permutation procedures, including notation and specific examples for small data sets. Thus, this 1936 article may well be the first example that detailed the method of calculating a permutation test using all possible arrangements of the observed data. It is interesting to note, however, that the work by Hotelling and Pabst became important in the discussion of distribution-free procedures involving rank data, but did not have a noticeable impact in the furthering of permutation tests.

2.12 Friedman and Analysis of Variance for Ranks

Trained as an economist, Milton Friedman became one of the most celebrated statisticians of his time. In addition to his contributions as an academic at the University of Chicago, he was also a public servant at the national level.

M. Friedman

(continued)

[47] This is an area of some controversy. Some researchers hold that, if and only if generalizing from a sample to a population, permutations are equally likely in controlled experimentation, but may not be equally likely in non-experimental research; see for example Zieffler, Harring, and Long [1493, pp. 132–134].

Milton Friedman graduated from Rutgers University in 1932 with an undergraduate degree in mathematics and economics, earned his M.A. degree from the University of Chicago in economics in 1933, and his Ph.D. in economics from Columbia University in 1946, where he worked with Harold Hotelling. During World War II, Friedman worked in Columbia's Statistical Research Group as a mathematical statistician (q.v. page 69). After the war, Friedman spent 1 year at the University of Minnesota where his good friend George Stigler was employed, but then accepted an appointment at the University of Chicago, where he taught for the next 30 years, while simultaneously maintaining a position with the National Bureau of Economic Research in New York City. Friedman was an academic who also spent much of his life in public service, but considered these part time activities, noting that his primary interest was his "scientific work" [487]. He was a member of President Ronald Reagan's Economic Policy Advisory Board and was awarded the Nobel Prize in Economic Sciences in 1976. Milton Friedman passed away on 16 November 2006 at the advanced age of 94 [483, 487].

Noting the contribution by Hotelling and Pabst on using rank data to overcome the assumption of normality in simple bivariate correlation, in 1937 Friedman outlined a similar procedure employing rank data in place of the ordinary analysis of variance [485].[48] If p denotes the number of ranks, Friedman utilized known results such as sums of natural integers, squared natural integers, and cubed natural integers from 1 to p given by $p(p+1)/2$, $p(p+1)(2p+1)/6$, and $p^2(p-1)^2/4$, respectively.

Friedman went on to show that the sampling distribution of the mean of ranks, where $\bar{r}_j$ denotes the mean rank of the jth of p columns, would have a mean value $\rho = (p+1)/2$ and a variance of $\sigma^2 = (p^2-1)/(12n)$, where n is the number of ranks averaged over the jth column. The hypothesis that the means come from a single homogeneous normal universe could then be tested by computing a statistic, χ_r^2, which Friedman noted tends to be distributed as the usual chi-squared distribution with $p-1$ degrees of freedom when the ranks are, in fact, random, i.e., when the factor tested has no influence [485, p. 676]. Friedman defined χ_r^2 as

$$\chi_r^2 = \frac{p-1}{p\,\sigma^2}\sum_{j=1}^{p}\left(\bar{r}_j-\rho\right)^2 = \frac{12n}{p(p+1)}\sum_{j=1}^{p}\left(\bar{r}_j-\frac{p+1}{2}\right)^2 ,$$

which for calculation purposes reduces to

[48] A clear and concise explanation of the Friedman analysis of variance for ranks test was given by Lincoln Moses in a 1952 publication on "Non-parametric statistics for psychological research" in *Psychological Bulletin* [1010].

$$\chi_r^2 = \frac{12}{np(p+1)} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} r_{ij} \right)^2 - 3n(p+1) \, ,$$

where r_{ij} denotes the rank in the ith of n rows and jth of p columns.

Friedman emphasized that the proposed method of ranks did not utilize all of the information provided by the observed data, as the method relied solely on the order of the variate and thus made no use of the quantitative magnitude of the variate. The consequences of that, he explained, were that (1) the method of ranks makes no assumption whatsoever as to the similarity of the distribution of the variate for the different rows, (2) the method of ranks does not provide for interaction because without quantitative measurements interaction is meaningless, and (3) the method of ranks is independent of the assumption of normality.

Friedman demonstrated that for $n = 2$, χ_r^2 tends to normality as p increases, and when n is large the discrete distribution of χ_r^2 approaches the continuous χ^2 distribution and the latter approaches normality as the degrees of freedom increases. For small samples, Friedman presented, in Tables V and VI in [485], the exact distribution of χ_r^2 in the case of $p = 3$ for $n = 2, \ldots, 9$ and in the case of $p = 4$, for $n = 2$, 3, and 4 [485, pp. 688–689]. Finally, returning to the work of Hotelling and Pabst, Friedman showed that the Spearman rank-order correlation coefficient investigated by Hotelling and Pabst was related to χ_r^2 when $n = 2$ as

$$\chi_r^2 = (p-1)(1-r') \, ,$$

where r' denotes the Spearman rank-order correlation coefficient. In 1997 Röhmel published an algorithm for computing the exact permutation distribution of the Friedman analysis of variance for ranks test [1186].

2.13 Welch's Randomized Blocks and Latin Squares

In 1937 B.L. Welch published an article in *Biometrika* that described permutation versions of randomized block and Latin square analysis of variance designs [1428]. He then compared the permutation versions of the two designs with the existing normal-theory versions.

B.L. Welch

Bernard Lewis Welch graduated with a degree in mathematics from Brasenose College, University of Oxford, in 1933. He then pursued a study of mathematical statistics at University College, London, where Pearson and Fisher had created a center for studies in statistical inference and biostatistics. Welch received an appointment to a Readership in Statistics in the University

(continued)

of Leeds, was appointed to the Chair in Statistics in 1968, and in the same year was appointed head of the newly created Department of Statistics. Bernard Lewis Welch suffered a stroke in June 1989 and died on 29 December of that same year; he was 78 years old [892].

In an article on randomized block and Latin square analysis of variance designs in *Biometrika* in 1937, Welch described Fisher's inference to an exact probability, referencing *The Design of Experiments*, and noted that although the calculations would be lengthy, the result would be a hypothesis test that was free of assumptions about the data [1428]. In this seminal article, Welch compared the normal-theory version of Fisher's variance-ratio z test with a permutation version in analyses of randomized block and Latin square designs.

Welch found it convenient to consider, instead of z, a monotonically increasing function of z given by

$$U = \frac{S_1}{S_0 + S_1} = \Big[(n-1)\exp(-2z) + 1\Big]^{-1} ,$$

where $S_1 = SS_{\text{Between}} = SS_{\text{Treatment}}$ and $S_0 = SS_{\text{Within}} = SS_{\text{Error}}$ in modern notation, although Jerzy Neyman had previously pointed out the advisability of considering the z-distribution directly [1033]. Like Eden and Yates in 1933 [379] and Pitman in 1937 [1129], Welch was able to reduce the amount of computation by considering only the variable portions of z. Welch explained that the convenience of U over z lies in the fact that in the permutation procedure $(S_0 + S_1)$ is constant, thus only the variation of $S_1 = SS_{\text{Between}}$ need be considered.

Utilizing the first two moments of the distribution of U, Welch analyzed a number of small published data sets in investigations of randomized block and Latin square designs. For randomized block designs, Welch found the expectations of differences and of mean squares based on permutations of the data generally to agree with those based on normal-theory methods. However, for Latin square designs Welch found that the permutation variance was considerably smaller than that of the normal-theory variance. Anticipating a debate that would appear and reappear in the permutation literature, Welch considered two possibilities for statistical inference. The first alternative considered a statistical inference about only the particular experimental data being analyzed; in Welch's case, a statistical inference only about the agricultural yields of a particular experimental field [1428, p. 48]. The second alternative considered the statistical inference drawn from the experimental data to a defined population, thus regarding the permutation distribution of z as a random sample from a set of similar distributions hypothetically obtained from other similar experiments [1428, p. 48].[49]

2.14 Egon Pearson on Randomization

E.S. Pearson, the son of Karl Pearson, had a distinguished career as a statistician in his own right. He collaborated extensively with Neyman and H.O. Hartley, among others, producing some of the most important and enduring statistical inference procedures of his time. His partnership with H.O. Hartley led to the two volume work on *Biometrika Tables for Statisticians* and his association with Jerzy Neyman led, of course, to the classical Neyman–Pearson approach to statistical inference, testing hypotheses, and confidence intervals.

E.S. Pearson

Egon Sharpe Pearson was the only son of Karl Pearson, who also had two daughters, and the two shared a deep interest in the history of probability and statistics [76]. E.S. Pearson was educated at Winchester College and Trinity College, University of Cambridge, but his education was interrupted by World War I. In 1920, Pearson was awarded a B.A. degree in mathematics after taking the Military Special Examination, set up by the British Government for those whose studies were delayed by the onset of the war. Pearson joined the Department of Applied Statistics, University College, London in 1920, where he attended lectures given by his father [814]. When Karl Pearson retired in 1933, the Department of Applied Statistics was divided into two departments. E.S. Pearson was appointed head of the Department of Applied Statistics and R.A. Fisher was appointed head of the Department of Eugenics.

Egon Pearson collaborated extensively with Jerzy Neyman (q.v. page 21) researching statistical inference [1035, 1036], an account of which is given by Pearson [1097], Reid [1160], and Lehmann [816, Chap. 3]. Pearson continued work begun by his father on editing the two volumes of *Tables for Statisticians and Biometricians*, collaborating with H.O. Hartley to compile and edit the tables that were eventually published as *Biometrika Tables for Statisticians, Volume I* in 1954 and *Biometrika Tables for Statisticians, Volume II* in 1972 [1101, 1102]. Pearson was elected Fellow of the Royal Society in 1966. Egon Sharpe Pearson F.R.S. died on 12 June 1980 at the age of 84.

[49] For a concise summary of the 1937 Welch paper, see a 2008 article by H.A. David on "The beginnings of randomization tests" in *The American Statistician* [326].

H.O. Hartley

Herman Otto Hartley (née Hirschfeld) fled Germany in 1934 shortly after completing his Ph.D. in mathematics at the University of Berlin to begin post-graduate work at the University of Cambridge. It was while in England that Hartley met E.S. Pearson at University College, London. In 1953, Hartley emigrated from England to the United States, joining the department of statistics at Iowa State University. In 1969, Hartley accepted a position as distinguished professor at Texas A&M University, and in 1979 Hartley was elected the 74th president of the American Statistical Association [321,1287]. Herman Otto Hartley passed away on 30 December 1980 in Durham, North Carolina, from complications following open heart surgery [321,1286,1287].

In 1937 E.S. Pearson referenced the Fisher text on *The Design of Experiments* in his consideration of randomizations in "Some aspects of the problem of randomization" [1093]. Pearson discussed the principle of randomization (i.e., permutation) and noted that most statistical tests used were developed on the assumption that the variables were normally distributed, but permutation tests, as developed by Fisher, were claimed to be independent of the assumption of normality. Pearson then asked "how far can tests be constructed which are completely independent of any assumption of normality?" [1093, p. 56].

Pearson provided concise summaries of several studies utilizing permutation methods, questioning whether the studies were truly independent of normality. The first study examined by Pearson was Fisher's investigation into Darwin's data on the heights of crossed- and self-fertilized *Zea mays* plants (q.v. page 62). Pearson noted that Fisher's study of the *Zea mays* plants found that 1,722 out of 32,768 possible values of the mean heights of plants were greater than the mean height of the observed plants, which was 20.933 in. (although the value given by Pearson of 1,722 appears to be a slight misprint) and that this was in no way unique. Pearson explained that Fisher could have used the geometric mean, for example, instead of the arithmetic mean and possibly found different results. The point being not that the geometric mean was a rational choice, but that "if variation is normal, a criterion based on the observed mean difference in samples [would] be most efficient in determining a real population difference" [1093, p. 58] and therefore using the arithmetic mean implied that the researcher believed a priori that the characteristics measured were likely to be normally distributed.

A second study examined by Pearson was Fisher's investigation into the coefficient of racial likeness [453]. As noted on page 65, Fisher considered measures of the statures of a random sample of $n = 100$ Frenchmen and $n = 100$ Englishmen to test the hypothesis that the mean heights of the sampled populations of Frenchmen and Englishmen were identical. Recall that Fisher conjectured writing the $2n$ measurements on cards, then shuffling the cards without regard to nationality.

Thus, it would be possible to divide the cards into two groups, each containing n cards, in $(2n)!/(n!)^2$ ways. The test statistic suggested was the difference between the means of the two groups. Again, Pearson questioned whether there was something fundamental about the form of the test "so it [could] be used as a standard against which to compare other more expeditious tests, such as Student's" [1093, p. 59].

Pearson continued with a hypothetical study based on two samples of seven observations each. The data for Samples 1 and 2 were: $\{45, 21, 69, 82, 79, 93, 34\}$ and $\{120, 122, 107, 127, 124, 41, 37\}$, respectively. Sample 1 had a mean of $\bar{x}_1 = 60.43$ and a midpoint, defined as the arithmetic average of the lowest and highest scores in the sample, of $m_1 = 57$; Sample 2 had a mean of $\bar{x}_2 = 96.86$ and a midpoint of $m_2 = 82$. He showed that after pooling the fourteen numbers, they could be divided into two groups of seven each in $(14!)/(7!)^2 = 3{,}432$ ways. Pearson found that the differences in means of the two samples had an equal or greater negative value than the observed mean difference of $\bar{x}_1 - \bar{x}_2 = 60.43 - 96.86 = -36.43$ in 126 out of 3,432 possible divisions, or 3.67 %. On the other hand, he found that the differences in midpoints of the two samples had an equal or greater negative value than the observed midpoint difference of $m_1 - m_2 = 57 - 82 = -25$ in 45 of the 3,432 divisions or, 1.31 %.

Pearson explained that random assignments of the 14 numbers into two groups of seven would give numerical values as large or larger than that observed to the difference in means on $2 \times 3.67 = 7.34\,\%$ of occasions, and numerical values as large or larger than that observed to the difference in midpoints on $2{\times}1.31 = 2.62\,\%$ of occasions. Pearson concluded that "applying this form of test to the midpoints, we would be more likely to suspect a difference in populations sampled than in applying the test to the means" [1093, p. 60]. Later in the article, Pearson confessed that he structured the data to favor the midpoints. Specifically, Pearson used Tippett's tables of uniform random numbers to draw the two samples from a rectangular distribution [1362]. Pearson showed that the standard error of the midpoint in samples of size n from a rectangular population with standard deviation σ_x was

$$\sigma_m = \sigma_x \sqrt{\frac{6}{(n+1)(n+2)}} = \sigma_x \sqrt{\frac{6}{(7+1)(7+2)}} = 0.289\,\sigma_x \ ,$$

while for the mean the standard error was considerably larger at

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{\sigma_x}{\sqrt{7}} = 0.378\,\sigma_x \ .$$

On this basis, Pearson argued "we should expect on theoretical grounds that the difference in sample midpoints, rather than in sample means, would be more efficient in detecting real differences" [1093, p. 61]. Pearson acknowledged that very few variables actually possess a rectangular distribution, but that he introduced these examples because they suggested that it is impossible to make a rational choice among alternative tests unless some information beyond that contained in the sample data is introduced. Pearson concluded the article with the acknowledgment that Fisher's randomization test was both exceedingly suggestive and often useful, but should be described as a valuable device rather than a fundamental principle.

As with Fisher, neither Welch nor Pearson fully explained the permutation technique. It was not until 1937 and 1938 that a series of articles by E.J.G. Pitman [1129–1131] explicitly discussed the permutation approach for statistical analysis. These three articles extended permutation methods to include data that were not amenable to ranking.

2.15 Pitman and Three Seminal Articles

E.J.G. Pitman, trained as a mathematician and isolated by distance from the centers of statistics in England due to his teaching duties at the University of Tasmania for 36 years, nonetheless contributed extensively to the early development of permutation methods. Some insight into Pitman the mathematician/statistician can be gleaned from a 1982 publication by Pitman titled "Reminiscences of a mathematician who strayed into statistics" in *The Making of Statisticians* edited by Joseph (Joe) Gani [1133].

E.J.G. Pitman

Edwin James George Pitman graduated from the University of Melbourne with a B.A. degree in mathematics in 1921, a B.Sc. degree in mathematics in 1922, and an M.A. degree in mathematics in 1923 [1458]. In 1926 Pitman was appointed Professor of Mathematics at the University of Tasmania, a position he held from 1926 to 1962. Like many contributors to statistical methods of this era, Pitman had no formal training in statistics, but was intrigued by the work of R.A. Fisher on statistical inference and randomization.

Pitman produced three formative papers on permutation methods in 1937 and 1938 [814, 1133, 1457]. In the introduction to the first paper on "Significance tests which may be applied to samples from any populations," Pitman first stated the object of the paper was to "devise valid tests of significance which involve no assumptions about the forms of the population sampled," and second, noted that the idea underlying permutation tests "seem[ed] to be implicit in all of Fisher's writings" [1129, p. 119]. Eugene Edgington, however, recounted that in 1986 Pitman expressed dissatisfaction with the introduction to his paper, writing "I [Pitman] was always dissatisfied with the sentence I wrote . . . I wanted to say I really was doing something new" (Pitman, quoted in Edgington [394, p. 18]). Edwin James George Pitman retired from the University of Tasmania in 1962 and died on 21 July 1993 at the age of 95.

2.15.1 Permutation Analysis of Two Samples

In the first of three seminal papers, Pitman demonstrated how researchers could devise valid tests of significance between two independent samples that made no assumptions about the distributions of the sampled populations. In addition, Pitman showed how precise limits could be determined for the difference between two independent means, again without making any assumptions about the populations from which the samples were obtained. An example will serve to illustrate Pitman's two-sample permutation test of significance. Consider two independent samples with m and n observations, respectively, and let $m \leq n$. Denote the observations in the first sample as $x_1,\ x_2, \ldots, x_m$ with mean $\bar{x}$, and denote the observations in the second sample as $y_1,\ y_2, \ldots, y_n$ with mean $\bar{y}$. Let the grand mean of the $m+n$ observations be given by

$$\bar{z} = \frac{m\bar{x} + n\bar{y}}{m+n}$$

and note that $\bar{z}$ is invariant over all

$$N = \binom{m+n}{m}$$

permutations of the $m+n$ observations with m and n held constant. Then

$$\bar{y} = \frac{1}{n}\Big[(m+n)\bar{z} - m\bar{x}\Big]$$

and the spread of the separation between $\bar{x}$ and $\bar{y}$ is given by

$$\begin{aligned} |\bar{x} - \bar{y}| &= \left|\bar{x} - \frac{1}{n}\Big[(m+n)\bar{z} - m\bar{x}\Big]\right| \\ &= \frac{m+n}{n}\left|\bar{x} - \bar{z}\right| \\ &= \left|\sum_{i=1}^{m} x_i - m\bar{z}\right| \frac{m+n}{mn} . \end{aligned}$$

Since m, n, and $\bar{z}$ are invariant over the permutations of the observed data, each arrangement of the observed data is a simple function of $\sum_{i=1}^{m} x_i$ for a one-sided probability value and $|\sum_{i=1}^{m} x_i - m\bar{z}|$ for a two-sided probability value; consequently, the computation required for each arrangement of the data is reduced considerably.

In contrast to contemporary permutation methods that compute the probability of an observed result as the proportion of simulated results as or more extreme than the observed result, Pitman devised a test of significance as follows. Let M be a fixed integer less than N and consider any particular mean difference denoted

Table 2.12 Eight groups of $m = 4$ with the largest values of $|\sum_{i=1}^{m} x_i - 68|$

Group	Groups of $m = 4$				$\sum_{i=1}^{m} x_i$	$\|\sum_{i=1}^{m} x_i - 68\|$
1	0	11	12	16	39	29
2	0	11	12	19	42	26
3	0	11	12	20	43	25
4	0	11	12	22	45	23
5	29	24	22	20	95	27
6	29	24	22	19	94	26
7	29	24	20	19	92	24
8	29	24	22	16	91	23

by R. If there are not more than M arrangements with a mean difference equal to or greater than that of R, the result is considered significant, and if there are M or more mean differences greater than that of R, the result is considered non-significant. As Pitman observed, in practice M is typically chosen to correspond with one of the usual working values, i.e., 5 or 1 %.

Pitman provided the following example, asking "Are the following samples significantly different?" $\{1.2, 2.3, 2.4, 3.2\}$ and $\{2.8, 3.1, 3.4, 3.6, 4.1\}$. To simplify calculation, Pitman subtracted 1.2 from each sample value, multiplied each difference by 10 to eliminate the decimal points, and re-arranged the nine values in order of magnitude, yielding $\{0, 11, 12, 16, 19, 20, 22, 24, 29\}$. He found the overall mean value to be $\bar{z} = 17$, so $m\bar{z} = 68$. Pitman explained that there were $N = (4 + 5)!/(4!\,5!) = 126$ of $m + n = 9$ values divided into samples of $m = 4$ and $n = 5$. The eight groups of $m = 4$ that gave the largest values of $|\sum_{i=1}^{m} x_i - 68|$ are listed in Table 2.12. Pitman observed that the third group of $\{0, 11, 12, 20\}$ gave the fifth largest value of $|\sum_{i=1}^{m} x_i - 68| = 25$ and was therefore significant at any level exceeding $5/126 = 0.0397$.

Importantly, Pitman noted that while only one test based on differences between two means was presented in this initial paper, the principle was applicable to all tests [1129, p. 119]. Pitman went on to mention that other tests of significance could be developed along the same lines, in particular an analysis of variance test, and commented that "the author hopes to deal with this in a further paper" [1129, p. 130].[50]

2.15.2 Permutation Analysis of Correlation

In the second of the three papers, Pitman began to fulfill his promise in the first paper and developed the permutation approach for the Pearson product-moment correlation coefficient "which makes no assumptions about the population

[50] H.A. David provides a concise summary of the 1937 Pitman paper in his 2008 article in *The American Statistician* on "The beginnings of randomization tests" [326].

sampled" [1130, p. 232]. Consider bivariate observations on n objects consisting of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, with means $\bar{x}$ and $\bar{y}$, respectively. Pitman showed that the observations of one set (x) may be paired with the observations of the other set (y) in $n!$ ways. Pitman's test of significance then paralleled the test of significance in the first paper. Pitman explained as follows. Let M be a fixed integer less than $N = n!$ and consider any particular pairing R. If there are not more than M pairings with a correlation coefficient equal to or greater than that of R in absolute value, then R is considered significant, and if there are M or more pairings with a correlation coefficient greater in absolute value than R, then R is considered non-significant.

Pitman summarized the results of his investigation by stating that the proposed test of significance for the correlation of a sample made no assumptions about the sampled population and concluded that some modification of the analysis of variance procedure would free it from its present assumptions, "but further discussion must be reserved for another paper" [1130, p. 232].

2.15.3 Permutation Analysis of Variance

True to form, Pitman followed up on this second promise in the third of his three papers, although this paper deviated somewhat from the presentations in the earlier two papers. In this third paper, Pitman proposed a permutation test for the analysis of variance "which involves no assumptions of normality" [1131, p. 335]. In this case, however, Pitman did not calculate a permutation test on actual data. Rather, Pitman detailed the mechanics and advantages of such a permutation test without carrying through the actual permutation analysis of experimental data, as he had in the previous two papers. Instead, Pitman noted that in the form of analysis of variance test discussed in the paper (randomized blocks) the observed numbers were not regarded as a sample from a larger population. Pitman went on to describe an experiment consisting of m batches, each batch composed of n individuals with the individuals of each batch subjected to n different treatments, and defined

$$W = \frac{SS_{\text{Treatment}}}{SS_{\text{Treatment}} + SS_{\text{Error}}} \,,$$

which is a monotonic increasing function of $SS_{\text{Treatment}}/SS_{\text{Error}}$.[51] Pitman explained that the problem of testing the null hypothesis that the treatments are equal is undertaken without making any assumptions. He went on to say that if the null hypothesis is true, then the observed value of W is the result of the chance allocation of the treatments to the individuals in the batches. He imagined repetitions of the same experiment with the same batches and the same individuals, but with different allocations of the treatments to the individuals in the various batches. Pitman also

[51] Pitman's use of $SS_{\text{Treatment}}$ and SS_{Error} is equivalent to SS_{Between} and SS_{Within}, respectively, as used by others.

noted that there were $N = (n!)^{m-1}$ ways in which the numbers may be grouped into n groups, so that W may take on N values, and that all values of W are equally-likely. However, Pitman stopped short of actually calculating a permutation test based on W. Instead he focused on deriving the first four moments of W and, based on the beta distribution, concluded that when both m and n are not too small, "the usual test may be safely applied" [1131, p. 335].[52]

2.16 Welch and the Correlation Ratio

In a 1938 article, "On tests for homogeneity," B.L. Welch (q.v. page 73) addressed tests of homogeneity for the correlation ratio, η^2. Assuming a set of k samples, Welch questioned whether they could reasonably be regarded as having all been drawn from the same population [1429]. Welch noted that η^2 depends on the observations having been drawn as random samples from an infinite hypothetical population and suggested that it may be better to consider the observations as samples from a limited population. Welch advocated calculating exact values on a limited population before moving into an examination of the moments of an infinite population [1429].

Welch explained that if there are N total observations with n_i observations in each treatment, $i = 1, \ldots, k$, then the N observations may be assigned to the k treatments in

$$\frac{N!}{n_1!\, n_2! \,\cdots\, n_k!}$$

ways and a discrete distribution of η^2 values may be constructed to which the observed value of η^2 may be referred [1429]. Welch continued with an example of an exact calculation and further concluded that if the variances of different samples were markedly different, normal-theory methods could badly underestimate significant differences that might exist. An exact permutation test, however, being free from the assumptions usually associated with asymptotic statistical tests, had no such limitation. Welch argued for the limited population approach on the grounds that it assumes nothing not obtained directly from the observed sample values.[53] However, Welch also noted that a limited population is only a mental construct. As an example, he pointed to a population of unemployed workers. This population definitely existed and could be sampled, but a population generated by shuffling the observed observations "does not correspond to anything concrete" other than the observed sample [1429, p. 154].

[52] The method of moments was first proposed by Karl Pearson in 1894 [1105].

[53] Today, this approach is termed "data-dependent" analysis.

2.17 Olds and Rank-Order Correlation

E.G. Olds, trained as a mathematician, nonetheless achieved substantial recognition in the fields of statistical assurance and quality control. In addition, Olds contributed to the growing literature on rank-order correlation methods begun by Spearman in 1904 [1300] and continued by Hotelling and Pabst in 1936 [653].

E.G. Olds

Edwin Glenn Olds graduated with a B.A. degree from Cornell University in 1918 and, at that point, went to Watkins (New York) High School as vice-principal and athletic coach, then became principal of Beeman Academy and the New Haven graded schools at New Haven, Vermont [284]. In 1923, Olds was appointed as instructor in mathematics at the Carnegie Institute of Technology [282].[54] Olds received his M.A. degree in mathematics from the University of Pittsburgh in 1925 [283] and his Ph.D. in mathematics from the University of Pittsburgh in 1931 [285], remaining at the Carnegie Institute of Technology for nearly 40 years [296]. Olds achieved considerable prominence in the fields of statistical assurance and quality control. Edwin Glenn Olds died following a heart attack on 10 October 1961 in his Pittsburgh home at the age of 61.

In 1938 Olds [1054], following up on the work by Hotelling and Pabst on rank-order correlation methods [653], calculated probability values up to $n = 10$ for Spearman's rank-order correlation coefficient [1300]. The probability values were based on the relative frequencies in the $n!$ permutations of one ranking against the other (q.v. page 4). The probability values for $n = 2, \ldots, 7$ were computed from exact frequencies, however those for $n = 8,\ 9$, and 10 were computed from Pearson type II curves.[55] Commenting on the difficulty of computing exact probability values, even for ranks, Olds echoed the frustration of many statisticians with the lack of computing power of the day, lamenting: "[f]or sums greater than 8 the [asymptotic] method becomes quite inviting" [1054, p. 141], and "[f]or n as small as 8, [an exact test] means the requirement of 42 formulas. It is fairly evident that these formulas will comprise polynomials ranging in degree from 0 to 41" [1054, p. 141]. Despite this, some 11 years later in 1949 Olds was able to extend the probability values for $n = 11,\ 12, \ldots, 30$, again employing Pearson type II curves [1055].

[54] In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University, which abuts the campus of the University of Pittsburgh. The Carnegie Institute of Technology is now the school of engineering at Carnegie Mellon University.

[55] There was an error in the denominator of the variance in the 1938 paper. It was first noticed by Scheffé in 1943 [1230] and corrected by Olds in 1949 [1055].

2.18 Kendall and Rank Correlation

M.G. Kendall is probably best remembered as the author of seminal books on rank-order correlation methods, advanced statistical methods, and a dictionary of statistical terms [729, 731, 734, 742]. However, he was also instrumental in the development and promotion of permutation statistical methods.

M.G. Kendall

Maurice George Kendall received his B.A. degree in mathematics from St. John's College, University of Cambridge, in 1929. In 1930, Kendall joined the British Civil Service in the Ministry of Agriculture, where he first became involved in statistical work. In 1949, Kendall accepted the second chair of statistics at the London School of Economics, which he held until 1961. Kendall spent the rest of his career in industry and in 1972 became Director of the World Fertility Study where he remained until 1980 when illness forced him to step down [1064]. Kendall is perhaps best remembered today for his revision of George Udny Yule's textbook *An Introduction to the Theory of Statistics* in 1937 [1482], first published in 1911 and continuing through 14 editions; Kendall's two volume work on *The Advanced Theory of Statistics*, with Volume I on "Distribution Theory" appearing in 1943 [729] and Volume II on "Inference and Relationship" in 1946 [731];[56] Kendall's definitive *Rank Correlation Methods*, first published in 1948; and Kendall's *Dictionary of Statistical Terms* with William R. Buckland, published in 1957 [742]. Kendall was knighted by Queen Elizabeth II in 1974 [73, 1328]. Sir Maurice George Kendall died on 29 March 1983 at the age of 75.

Kendall incorporated exact probability values utilizing the "entire universe" of permutations in the construction of τ, a new measure of rank-order correlation in 1938 [728].[57] The new measure of rank correlation was based on the difference between the sums of the concordant and discordant pairs of observations. The actual score for any given ranking of the data was denoted as Σ by Kendall. For example, consider the data of two sets (A and B) of ten ranks in Fig. 2.3. There are $n(n-1)/2 = 10(10-1)/2 = 45$ possible pairs, divisible into concordant and

[56] While *The Advanced Theory of Statistics* began as a two-volume work, in 1966 Alan Stuart joined with Maurice Kendall and *The Advanced Theory* was rewritten in three volumes. Keith Ord joined in the early eighties and a new volume on Bayesian Inference was published in 1994. More recently, Steven Arnold was invited to join with Keith Ord.

[57] As Kendall explained in a later publication, the coefficient τ was considered earlier by Greiner [554] and Esscher [414] as a method of estimating correlations in a normal population, and was rediscovered by Kendall [728] who considered it purely as a measure of rank-order correlation [734].

Fig. 2.3 Sets A and B of ten ranks each

A:	1	2	3	4	5	6	7	8	9	10
B:	4	7	2	10	3	6	8	1	5	9

Fig. 2.4 Successive arrays of Σ values as delineated by Kendall [728]

<table>
<tr><th colspan="11">Arrays of Σ values for $n = 1, \ldots, 5$</th></tr>
<tr><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>1</td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>1</td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td>1</td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>1</td><td>2</td><td>2</td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>1</td><td>2</td><td>2</td><td>1</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td>1</td><td>2</td><td>2</td><td>1</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td>1</td><td>2</td><td>2</td><td>1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>1</td><td>3</td><td>5</td><td>6</td><td>5</td><td>3</td><td>1</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td>1</td><td>3</td><td>5</td><td>6</td><td>5</td><td>3</td><td>1</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td>1</td><td>3</td><td>5</td><td>6</td><td>5</td><td>3</td><td>1</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td>1</td><td>3</td><td>5</td><td>6</td><td>5</td><td>3</td><td>1</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td>1</td><td>3</td><td>5</td><td>6</td><td>5</td><td>3</td><td>1</td></tr>
<tr><td>1</td><td>4</td><td>9</td><td>15</td><td>20</td><td>22</td><td>20</td><td>15</td><td>9</td><td>4</td><td>1</td></tr>
</table>

discordant pairs of observations. A concordant pair has the same order and sign and a discordant pair has a different order and sign. For example, the first pair, starting from the left, is $A = \{1, 2\}$ and $B = \{4, 7\}$. Since $1 - 2 = -1$ and $4 - 7 = -3$, the first pair is concordant as both signs are negative. The second pair is $A = \{1, 3\}$ and $B = \{4, 2\}$ and since $1-3 = -2$ and $4-2 = +2$, the second pair is discordant as the signs do not agree, with one being negative and the other positive. The last pair is $A = \{9, 10\}$ and $B = \{5, 9\}$ and since $9 - 10 = -1$ and $5 - 9 = -4$, the last pair is concordant as the signs agree. For these data, the number of concordant pairs is 25 and the number of discordant pairs is 20. Thus, $\Sigma = 25 - 20 = +5$ for these data.

Kendall considered the entire universe of values of Σ obtained from the observed rankings $1, 2, \ldots, n$ and the $n!$ possible permutations of the n ranks (q.v. page 4). A clever recursive procedure permitted the calculation of the frequency array of Σ, yielding a figurate triangle similar to Pascal's triangle.[58]

As Kendall explained, the successive arrays of Σ were constituted by the process illustrated in Fig. 2.4. For each row, to find the array for $(n + 1)$, write down the nth array $(n + 1)$ times, one under the other and moving one place to the right each

[58] A recursive process is one in which items are defined in terms of items of similar kind. Using a recurrence relation, a class of items can be constructed from a few initial values (a base) and a small number of relationships (rules). For example, given the base, $F_0 = 0$ and $F_1 = F_2 = 1$, the Fibonacci series $\{0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots\}$ can be constructed by the recursive rule $F_n = F_{n-1} + F_{n-2}$ for $n > 2$.

Fig. 2.5 Figurate triangle for values of Σ with $n = 1, \ldots, 5$

n	Figurate triangle										
1	1										
2	1	1									
3	1	2	2	1							
4	1	3	5	6	5	3	1				
5	1	4	9	15	20	22	20	15	9	4	1

time, and then sum the $(n + 1)$ arrays. The process may be condensed by forming a figurate triangle as in Fig. 2.5. Here, a number in the nth row is the sum of the number immediately above it and the $n - 1$ (or fewer) numbers to the immediate left of that number.

Consider row $n = 5$ in the figurate triangle in Fig. 2.5 where the value of 4 in the second position from the left in row 5 is the sum of the number above it (3) in row 4 and all the numbers to the left of 3 in row 4 (1), since there are fewer than $n - 1 = 4$ numbers to the left of 3; the value of 9 in the third position from the left in row 5 is the sum of the number above it (5) in row 4 and all the numbers to the left of 5 in row 4 (3 and 1), since there are fewer than $n - 1 = 4$ numbers to the left of 3; the value of 15 in the fourth position from the left in row 5 is the sum of the number above it (6) in row 4 and all the numbers to the left of 3 in row 4 (5, 3, and 1), since there are fewer than $n - 1 = 4$ numbers to the left of 6; the value of 20 in the fifth position from the left in row 5 is the sum of the number above it (5) in row 4 and all the numbers to the left of 5 in row 4 (6, 5, 3, and 1), since there are $n - 1 = 4$ numbers to the left of 5; and the value of 22 in the sixth position from the left in row 5 is the sum of the number above it (3) in row 4 and the $n - 1 = 4$ numbers to the left of 3 in row 4 (5, 6, 5, and 3), since there are more than $n - 1 = 4$ numbers to the left of 3. The terms to the right of the last number are filled in from the left, as each array is symmetrical. A check is provided by the fact that the total in the nth row is equal to $n!$. Utilizing this technique, Kendall was able to construct a table of the distribution of Σ for values of n from 1 to 10 [728, p. 88].

This accomplishment was further extended in a 1939 publication in which Kendall and Bernard Babington Smith considered "The problem of m rankings," developing the well-known coefficient of concordance [739].[59,60] Let n and m denote the number of ranks and the number of judges, respectively, then Kendall and Babington Smith defined the coefficient of concordance, W, as

$$W = \frac{12S}{m^2(n^3 - n)} \,,$$

[59] A correction was proffered by J.A. van der Heiden in 1952 for observers who declined to express a preference between a pair of objects [1390].

[60] The coefficient of concordance was independently developed by W. Allen Wallis in 1939, which he termed the "correlation ratio for ranked data" [1411].

where S is the observed sum of squares of the deviations of sums of ranks from the mean value $m(n+1)/2$. W is simply related to the average of the $\binom{m}{2}$ Spearman rank-order correlation coefficients between pairs of m rankings. Kendall and Babington Smith showed that the average Spearman rank-order correlation, ρ_{av}, is given by

$$\rho_{av} = \frac{mW - 1}{m - 1}$$

and pointed out that ρ_{av} is simply the intraclass correlation coefficient, r_I, for the m sets of ranks. The coefficient of concordance is also equivalent to the Friedman two-way analysis of variance for ranks, as noted by I.R. Savage in 1957 [1224, p. 335].

Since $m^2(n^3 - n)$ is invariant over permutation of the observed data, Kendall and Babington Smith showed that to test whether an observed value of S is statistically significant it is necessary to consider the distribution of S by permuting the n ranks in all possible ways. Letting one of the m sets of ranks be fixed, then there are $(n!)^{m-1}$ possible values of S. Based on this permutation procedure, Kendall and Babington Smith created four tables that provided exact probability values for $n = 3$ and $m = 2, \ldots, 10$, $n = 4$ and $m = 2, \ldots, 6$, and $n = 5$ and $m = 3$.

In the same year, 1939, Kendall, Kendall, and Babington Smith utilized permutation methods in a discussion of the distribution of Spearman's coefficient of rank-order correlation, ρ_s, introduced by Spearman in 1904 [1300] and given by

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \; ,$$

where $d_i = X_i - Y_i$ and X_i and Y_i, $i = 1, \ldots, n$, are the permutation sequences of the natural integers from 1 to n [746]. Kendall, Kendall, and Babington Smith observed that to judge the significance of a value of ρ_s it is necessary to consider the distribution of values obtained from the observed ranks with all other permutations of the numbers from 1 to n and further noted that in practice it is generally more convenient to consider the distribution of $\sum_{i=1}^{n} d_i^2$ [746, p. 251]. They remarked that distributions for small values of n obtained by Hotelling and Pabst [653] deviated considerably from normality and that Hotelling and Pabst proved that as $n \to \infty$ the distribution of ρ_s tends to normality. They went on to mention that ρ_s is mainly of service when $10 \leq n \leq 30$ and stated that "it is the aim of the present paper to throw some light on this crepuscular territory" [746, p. 252]. Finally, Kendall, Kendall, and Babington Smith gave explicit values up to and including $n = 8$ with some experimental distributions for $n = 10$ and $n = 20$. The distributions for n up to 8 were exact and the distributions for $n = 10$ and $n = 20$ were based on a random sample of 2,000 permutations [746, pp. 261–267].

2.19 McCarthy and Randomized Blocks

M.D. McCarthy, trained as a statistician, was both an accomplished academic and an able administrator, ultimately serving for 11 years as president of University College, Cork, in Ireland. McCarthy urged researchers to first use a permutation test as an approximation to a normal-theory test, then apply the normal-theory test.

M.D. McCarthy

M. Donal McCarthy received most of his advanced education at University College, Cork, earning a B.A. degree in mathematics and mathematical physics in 1928, an M.Sc. degree in mathematical science in 1934, and a Ph.D. in statistics in 1938. He was an academic until he was appointed Director of the Central Statistics Office, Ireland, on the resignation of R.C. Geary, serving from 1957 to 1966. From 1967 to 1978 he served as President of University College, Cork. M. Donal McCarthy died on 31 January 1980 at the age of 71 [910].

In 1939 McCarthy [911] also argued for the use of a permutation test as a first approximation before considering the data via an asymptotic distribution, citing earlier works by Fisher in 1935 [451] and 1936 [453] as well as by Welch in 1938 [1429]. McCarthy explained that in certain experiments, especially those in the physical and chemical sciences, it is possible for a researcher to repeat an experiment over and over. The repetition provides a series of observations of the "true value," subject only to random errors. However, in the biological and social sciences it is nearly impossible to repeat an experiment under the same essential conditions. McCarthy addressed the problem of analyzing data from a randomized blocks experiment and utilized Fisher's variance-ratio z statistic (q.v. page 33). He concluded that the use of the z statistic is theoretically justifiable only when the variations within each block are negligible, and suggested a permutation test on the yields from a single block as a first approximation.

2.20 Computing and Calculators

The binary (base 2) system is the foundation of virtually all modern computer architecture. Although the full documentation of the binary system is usually attributed to the German philosopher and mathematician Gottfried Leibniz[61] in his 1703 article on "Explication de l'arithmétique binaire" (Explanation of binary arithmetic)

[61] Also spelled Leibnitz.

[508, pp. 223–227], priority should probably be given to the English mathematician and astronomer Thomas Harriot[62] [357, 1047, 1266].

T. Harriot

Thomas Harriot, born circa 1560 in Oxfordshire, England, was an astronomer, mathematician, ethnographer, translator, and the founder of the English school of algebra [1047]. He graduated from St. Mary's Hall, University of Oxford, in 1580 and immediately moved to London. In 1583 Harriot entered Sir Walter Raleigh's service as a cartographer, navigational instructor to Raleigh's seamen, Raleigh's accountant, and designer of expeditionary ships. He sailed with Raleigh to Virginia in 1585–1586 and most probably accompanied Raleigh on his expedition to Roanoke Island off the coast of North Carolina in 1584. Harriot translated the Carolina Algonquin language from two native Americans, Wanchese and Manteo, who had been brought back to England by Raleigh in 1584 [586].

In the 1590s Harriot moved from working with Raleigh to an association with Henry Percy, the 9th Earl of Northumberland. The Earl introduced him to a circle of scholars, gave him property in the form of a former Abbey, and provided him with a handsome pension and a house on Northumberland's estate of Syon House, west of London on the Thames River near Kew, that Harriot used as both a residence and a scientific laboratory. Harriot is best known for his work on algebra, introducing a simplified notation and working with equations of higher degrees [1392]. Harriot published only one book in his lifetime, leaving unpublished some 7,000 pages of hand-written manuscripts that have slowly come into the mainstream of historical record over the past three centuries. The book, published in 1588, was an abstract of his extensive *Chronicle* (now lost) as *A Briefe and True Report of the New Found Land of Virginia*—the first book in English about the New World, detailing the flora, fauna, and land resources of Virginia [587].

As described on the website of the Thomas Harriot College of Arts and Sciences, Harriot was a man of both intellect and action, described by a contemporary as, "[t]he master of all essential and true knowledge." He played many roles as an adventurer, anthropologist, astronomer, author, cartographer, ethnographer, explorer, geographer, historian, linguist, mathematician, naturalist, navigator, oceanographer, philosopher, planner, scientist, surveyor, versifier, and teacher [586]. The sweeping breadth of Harriot's life story is well told in John W. Shirley's book *Thomas Harriot: A Biography* [1267]. In addition, the Thomas Harriot College of Arts and Sciences at East Carolina State University in Greenville, North Carolina, maintains a list of Internet

(continued)

[62] Also spelled Hariot, Harriott, or Heriot.

web-based sources on Thomas Harriot and his times [1265]. Thomas Harriot died on 2 July 1621 in London and was buried in St. Christopher le Stocks, which was destroyed in the Great Fire of London in 1666 and is presently the site of the Bank of England.

G.W. Leibniz

Gottfried Wilhelm von Leibniz was born on 1 July 1646 in Leipzig, Saxony, although some sources put the date of birth as 21 June 1646 using the Julian calendar. In 1661 Liebniz began his university education at the University of Leipzig. After earning his B.A. from Leipzig in December 1662, he continued his studies at the University of Altdorf, earning a Doctorate of Law in 1667. While at Altdorf, Leibniz published his *Dissertation de arte combinatoria* (Dissertation on the Art of Combinations) in 1661 at the age of 20. In 1672 the Elector of Mainz, Johann Philipp von Schönborn, sent Leibniz on a diplomatic mission to Paris, then the center of learning and science. He remained in Paris for 4 years, meeting with many of the major figures of the intellectual world. In addition, he was given access to the unpublished manuscripts of both René Descartes and Blaise Pascal. It was upon reading these manuscripts that he began to conceive of the differential calculus and his eventual work on infinite series [842].

In 1673 Leibniz traveled to London to present a prototype of his Stepped Reckoner calculating machine to the Royal Society. In 1676 Leibniz was appointed to the position of Privy Counselor of Justice to the Duke of Hanover, serving three consecutive rulers of the House of Brunswick in Hanover as historian, political advisor, and as librarian of the ducal library. Leibniz is considered by modern scholars as the most important logician between Aristotle and the year 1847, when George Boole and Augustus De Morgan published separate books on modern formal logic. In addition, Leibniz made important discoveries in mathematics, physics, geology, paleontology, psychology, and sociology. Leibniz also wrote extensively on politics, law, ethics, theology, history, and philosophy [819].

Today. Leibniz is best remembered, along with Sir Isaac Newton, for the invention of infinitesimal calculus. He introduced many of the notations used today, including the integral sign, $\int$, and the d used for differentials. Gottfried Wilhelm von Leibniz died in Hanover on 14 November 1716.

While Leibniz invented the Stepped Reckoner, a decimal (non-binary) calculator that could add (subtract) an 8 digit number to (from) a 16 digit number, multiply two 8 digit numbers together by repeated addition, or divide a 16 digit number by an 8 digit divisor by repeated subtraction, computing by machine had its beginnings with the work of Charles Babbage, variously referred to as the "Grandfather" or the "Patron Saint" of computing. Sometime around 1821, Babbage had the idea to develop mechanical computation. Babbage was frustrated with the many errors in tables used for calculating complex equations, some of which had persisted for hundreds of years. The errors were largely due to the fact that the tables were copied by hand and further transcribed to plates for printing. This led Babbage to develop a mechanical device to calculate and print new tables; the device was called the Difference Engine as it was designed for calculating polynomials of higher orders using the method of differences [1336]. The Difference Engine was never finished by Babbage, but was finally constructed in 1991 and presently resides in the London Science Museum.[63]

C. Babbage

Charles Babbage was born in London on 26 December 1791, the son of a London banker. He attended Trinity College, University of Cambridge, in 1810 but was disappointed in the level of mathematical instruction available at the time at Trinity. In 1812 he transferred to Peterhouse College, University of Cambridge, graduating in 1814. In 1817 Babbage received an M.A. degree from Cambridge. In his twenties, Babbage worked as a mathematician and was a founder of the Analytical Society along with George Peacock, John Herschel, Michael Slegg, Edward Bromhead, Alexander D'Arblay, Edward Ryan, Frederick Maule, and others. In 1821 Babbage invented the Difference Engine to compile mathematical tables [106, 1290]. From 1828 to 1839 Babbage occupied the Lucasian Chair of Mathematics[64] at the University of Cambridge—Isaac Newton's former position and one of the most prestigious professorships at Cambridge—and played an important role in the establishment of the Astronomical Society with mathematician and astronomer John Frederick William Herschel, the London Statistical Society in 1834 (later, in 1887, the Royal Statistical Society) and the British Association for the Advancement of Science (BAAS) in 1831 [1027]. In 1856 he conceived of a general symbol manipulator, the Analytical Engine.

As an interesting aside, in 1833, at a meeting of the British Association for the Advancement of Science (now, the British Science Association) the poet Samuel Taylor Coleridge raised the question as to what name to give to professional experts in various scientific disciplines: an umbrella term that

(continued)

[63] Actually, the model in the London Science Museum is of Difference Engine Number 2, designed by Babbage between 1846 and 1849 [1290, pp. 290–291].

[64] In a wonderful little book on the history of British science during the nineteenth century, Laura Snyder noted that while Lucasian Professor of Mathematics at the University of Cambridge from 1828 to 1839, Charles Babbage never delivered a single lecture [1290, p. 130].

would include anatomists, astronomers, biologists, chemists, and others. The word "scientist" was suggested by William Whewell, a mineralogist, historian of science, and future master of Trinity College, and thus was coined the term "scientist" [1175, p. 8].

Babbage published some eighty volumes in his lifetime and was elected Fellow of the Royal Society in 1816. Among other accomplishments, Babbage published a table of logarithms from 1 to 108,000 in 1827 and invented the cow-catcher, the dynamometer, the standard railroad gauge, and occulting lights for lighthouses. Charles Babbage F.R.S. passed away at home in London on 18 October 1871 at the age of 79 [672, 1447].

In a well-known story, the textile industry served as the stimulus for Babbage to provide instructions to the Difference Engine. On 30 June 1836 Babbage conceived the idea of using punch cards like those devised by Joseph-Marie Jacquard in 1801 to produce patterns in weaving looms. These were similar in both form and function to those used by Herman Hollerith in 1884 for his electric punch-card tabulator. Babbage devised a system using four different types of punch cards, each about the height and width of a modern-day brick. Operation cards instructed the engine to add subtract, multiply, or divide; variable cards instructed the engine from where to retrieve the number and where to store the result; combinatorial cards instructed the engine to repeat a set of instructions a specified number of times; and number cards were used to save the results [1290, p. 215].

The Jacquard Loom

The Jacquard loom used a series of cards with tiny holes to dictate the raising and lowering of the warp threads. The warp threads are the longitudinal threads and the weft threads are the lateral threads. In the weaving process, the warp threads are raised and lowered as the weft threads are passed through to create the textile. Rods were linked to wire hooks, each of which could lift one of the warp threads. The cards were pressed up against the ends of the rods. When a rod coincided with a hole in the card, the rod passed through the hole and no action was taken with the thread. On the other hand, if no hole coincided with a rod, then the card pressed against the rod and this activated the wire hook that lifted the warp thread, allowing the shuttle carrying the weft to pass underneath the warp thread [1290, p. 214–215]. The arrangement of the holes determined the pattern of the weave. The Jacquard method, for intricate weaving, could require as many as 20,000 punched cards with 1,000 holes per card.

Fig. 2.6 Example of generating successive values for $f(x) = 3x^2 - 2x + 5$ using the method of differences

Column			
1	2	3	4
x	$f(x)$	Δ_1	Δ_2
0	5		
		1	
1	6		6
		7	
2	13		6
		13	
3	26		6
		19	
4	45		**6**
		25	
5	**70**		**6**
		31	
6	**101**		

2.20.1 The Method of Differences

The method of differences defines a process for calculating complex polynomial expressions using only addition—no multiplication or division—thereby making it highly amenable to machine calculation. To illustrate the method of differences, consider a second degree polynomial $f(x) = 3x^2 - 2x + 5$. Figure 2.6 demonstrates how the method of differences works. Column 1 in Fig. 2.6 lists possible values of x from 0 to 4 in Roman typeface, where 4 is the order of the polynomial plus 2. Column 2 evaluates the polynomial expression $f(x) = 3x^2 - 2x + 5$. Column 3 lists difference values for $\Delta_1 = f(x + 1) - f(x)$ obtained from Column 2, commonly called first-order differences. Column 4 lists the second-order differences $\Delta_2 = \Delta_1(x + 1) - \Delta_1(x)$ that yield a common value of 6. For any polynomial of order n, Column $n + 2$ will be a constant.

Once stasis has been reached in Column $n + 2$, additional values of x can be evaluated by simple addition by reversing the process. Add an additional value of the constant **6** to Column 4 (shown in bold typeface); then add that value (**6**) to the last value in Column 3 (**6** + 19 = **25**); add that value (**25**) to the last value in Column 2 (**25** + 45 = **70**); and finally increment Column 1 by 1 (4 + 1 = **5**). For the next step add another value of **6** to Column 4; add that **6** to the last value in Column 3 (**6** + **25** = **31**); add the **31** to the last value in Column 2 (**31** + **70** = **101**); and increment Column 1 by 1 (**5** + 1 = **6**). The process can be continued indefinitely.

2.20.2 Statistical Computing in the 1920s and 1930s

Permutation methods, by their very nature, incorporate computationally-intensive procedures and it would be imprudent not to mention the tabulating procedures of the 1920s and 1930s. Fisher had purchased a Millionaire calculator soon after he arrived at the Rothamsted Experimental Station in 1919.[65] While addition, subtraction, and multiplication were easy to implement on the Millionaire, division was not, and hand-written tables of reciprocals were attached to the lid of the Millionaire to ease the problem [1027].[66] Fisher's original Millionaire was still in the office of Frank Yates at Rothamsted in 1974.[67] Karl Pearson relied on his beloved Brunsviga calculators at the Galton Biometric Laboratory, which were noisy, limited, but very robust machines. Division was done by repeated subtraction until a bell rang to indicate passage through zero [1027]. Toward the end of his life in 1936, Pearson was still using a vintage Brunsviga that dated from the turn of the century and Maurice Kendall was using a Brunsviga in 1965 that he had inherited from Udny Yule [1164, p. 18]. Commenting on the use of mechanical desk calculators between 1945 and 1969, M.G. Kendall wrote:

> [p]ractical statistics was conditioned by what such a machine — or in a few favored cases, a battery of such machines — could accomplish. In consequence theoretical advance was held back, not so much by the shortage of ideas or even of capable men to explore them as by the technological impossibility of performing the necessary calculations. The Golden Age of theoretical statistics was also the age of the desk computer. Perhaps this was not a net disadvantage. It generated, like all situations of scarcity, some very resourceful shortcuts, economies, and what are known unkindly and unfairly as quick and dirty methods. But it was undoubtedly still a barrier [738, p. 204].

Statistical computing in the United States in the 1920s was concentrated in modest statistical laboratories scattered around the country and employed small mechanical desk calculators such as those manufactured by the Burroughs, Victor, Monroe, Marchant, or Sundstrand companies [557]. Grier provides an excellent historical summary of the development of statistical laboratories in the United States in the 1920s and 1930s [557] and Redin provides a brief but comprehensive history of the development of mechanical calculators in this period [1158]. Most of these research laboratories were small ad hoc university organizations and many were nothing more than a single faculty member arranging to use the university tabulating machines during off hours [557]. The largest of these laboratories were substantial organizations funded by small foundations or by private individuals. One of the first of these statistical computing laboratories was founded at the University of Michigan by James Glover, a professor of mathematics, under whom George Snedecor studied. Interest in statistical computing became a popular field of study

[65] The Millionaire calculator was the first commercial calculator that could perform direct multiplication. It was in production from 1893 to 1935.

[66] For Fisher's first major publication in 1921 on "Studies in crop variation, I," Fisher produced 15 tables [445]. At approximately 1 min for each large multiplication or division problem, it has been estimated that Fisher spent 185 h using the Millionaire to produce each of the 15 tables [618, p. 4].

[67] For pictures of the Millionaire calculator and Frank Yates using the Millionaire, see a 2012 article by Gavin Ross in *Significance* [1196]. Also, there is a YouTube video of a Millionaire calculator calculating the surface of a circle with diameter 3.18311 at http://www.youtube.com/watch?v=r9Nnl-u-Xf8.

during the 1930s, as research laboratories acquired the early punch-card tabulator, first developed by Herman Hollerith for the 1890 census [557]. A picture of the Hollerith 1890 census tabulator can be viewed at a website on computing history constructed by Frank da Cruz [308].

H. Hollerith

Herman Hollerith, often called "the father of automatic computation," graduated from Columbia University with an Engineer of Mines (EM) degree in 1879 and then worked for the U.S. Bureau of the Census on the 1880 census. Hollerith quickly determined that if numbers could be punched as holes into specific locations on cards, such as used to produce patterns in a Jacquard weaving loom, then the punched cards could be sorted and counted electromechanically. The punched cards were especially designed by Hollerith, having one corner cut off diagonally to protect against the possibility of upside-down or backwards cards and each punched card was constructed to be exactly 3.25 in. wide by 7.375 in. long, designed to be the same size as the 1887 U.S. paper currency because Hollerith used Treasury Department containers as card boxes. The actual size of the United States currency in 1887 was approximately 3.125 in. wide by 7.4218 in. long (79 mm × 189 mm), with modern currency introduced in 1929 measuring 2.61 in. wide by 6.14 in. long (66.3 mm × 156 mm).

Hollerith submitted a description of this system, *An Electric Tabulating System* [640, 641], to Columbia University as his doctoral thesis and was awarded a Ph.D. from Columbia University in 1890. There has always been a suspicion that this was an honorary degree, but it has recently been definitively established that the degree was not an honorary degree and was awarded by the Board of Trustees granting Hollerith "the degree of Doctor of Philosophy upon the work which he has performed" [308].

Hollerith went on to invent a sorter and tabulating machine for the punched cards, as well as the first automatic card-feed mechanism and the first key punch. On 8 January 1889 Hollerith was issued U.S. Patent 395,782 for automation of the census. It should be noted that the 1880 census with 50 million people to be counted took over 7 years to tabulate, while the 1890 census with over 62 million people took less than a year using the tabulating equipment of Hollerith (different sources give different numbers for the 1890 census, ranging from 6 weeks to 3 years) [308].

In 1896 Hollerith started his own business, founding the Tabulating Machine Company. Most of the major census bureaus in Russia, Austria, Canada, France Norway, Puerto Rico, Cuba, and the Philippines leased his tabulating equipment and purchased his cards, as did many insurance companies. In 1911 financier Charles R. Flint arranged the merger of the

(continued)

Tabulating Machine Company, the International Time Recording Company, and the Computing Scale Company to form the Computing Tabulating Recording Corporation (CTR). In 1914 Flint recruited Thomas J. Watson from the National Cash Register (NCR) Company to lead the new company. In 1924 CTR was renamed the International Business Machines Corporation (IBM). Herman Hollerith passed away on 17 November 1929 in Washington, DC.

In the absence at that time of government granting agencies such as the National Science Foundation (NSF) and the National Institutes of Health (NIH), it fell to the United States Department of Agriculture (USDA) to establish the largest of the early statistical laboratories: the Statistical Laboratory at Iowa State College (now, Iowa State University) under the direction of George W. Snedecor in 1933 (q.v. page 35).[68] Snedecor previously had been trained by James Glover in the Statistical Laboratory at the University of Michigan.

The Graduate College of Iowa State College was always alert for opportunities to invite outstanding scientists to visit and give lectures on their recent work. This helped keep the local staff abreast of promising developments at other research centers. Largely due to Dean R.E. Buchanan of the Graduate College and Professor E.W. Lindstrom of the Department of Genetics, it was the regular custom through the 1930s and 1940s to invite an outstanding scientist as a Visiting Professor for 6 weeks each summer. The Graduate College provided the expenses and honorarium of the visiting scientist [859]. In 1931 and 1936 Snedecor invited R.A. Fisher to visit the Department of Statistics at Iowa State College for the summer. Fisher's lodging was a room on the second floor of the Kappa Sigma (KΣ) fraternity house several blocks from the Iowa State campus. To combat the summer heat in Iowa, Fisher would put the sheets from his bed into the refrigerator for the day, then remake his bed every evening [576].[69]

While Fisher was at Iowa State College in 1936, the college awarded him an honorary D.Sc. degree, his first of many.[70] Over the two summers, Fisher met and worked with about 50 researchers eager to learn his methods of analysis. One of these researchers was Henry Agard Wallace, who later left Iowa State College to become Secretary of Agriculture.[71] As Secretary, Wallace devised and prepared

[68] Iowa Agricultural College and Model Farm was established in 1858 and changed its name to Iowa State University of Science and Technology in 1959, although it is commonly known as Iowa State University.

[69] For more interesting stories about Fisher, see a 2012 article in *Significance* by A.E.W. Edwards and W.F. Bodmer [401].

[70] Interestingly, the Statistical Laboratory at Iowa State College initiated four o'clock afternoon tea while Fisher was there in the summer of 1936 [57, 576].

[71] Henry A. Wallace served as Secretary of Agriculture from 1933 to 1940. When John Nance Garner broke with then President Franklin Delano Roosevelt in 1940, Roosevelt designated Wallace to run as his Vice-President. Wallace served as Vice President from 1941 to 1945 when

the Agricultural Adjustment Act, which required the Department of Agriculture to undertake large studies of major farm products. Thus, the Agricultural Adjustment Act of 1933 was a boon to the Statistical Laboratory at Iowa State College.[72] Coincidentally, the first statistical computing laboratory to use a punched-card tabulator was not a university laboratory, but the computing laboratory of the Bureau of Agricultural Economics, a division of the Department of Agriculture, which started using punched cards in 1900 [557].

2.21 Looking Ahead

A number of notable threads of inquiry were established during the period 1920 to 1939 that were destined to become important in the later development of permutation methods.

1. There was widespread recognition of the computational difficulties inherent in constructing permutation tests by hand, with several researchers bemoaning the restriction of permutation methods to small samples. For example, Hotelling and Pabst were forced to limit construction of their exact tables for Spearman's rank-order correlation coefficient to small samples of $n = 2, 3$, and 4, noting that for larger samples the calculation of exact probability values would be very laborious [653, p. 35]. Like Hotelling and Pabst, Olds calculated probability values up to $n = 10$ for Spearman's rank-order correlation coefficient, but only the probability values for $n = 2, \ldots, 7$ were calculated exactly; those for $n = 8$, 9, and 10 were approximated by Pearson type II curves [1054]. In like manner, Kendall, utilizing a recursion procedure, was able to provide exact probability values for the τ measure of rank-order correlation, but only up to $n = 10$ [728].
2. Throughout the period 1920–1939 there was general acceptance that permutation tests were data-dependent, relying solely on the information contained in the observed sample without any reference to the population from which the sample had been drawn. Thus, permutation tests were considered to be distribution-free and not restricted by any assumptions about a population, such as normality. For example Frank Yates, commenting on the experiment on Yeoman II wheat shoots conducted by Thomas Eden and himself, concluded that the need for the postulation of any parent population from which the observed

Roosevelt jettisoned Wallace in favor of Harry S. Truman, who succeeded Roosevelt upon his death on 12 April 1945 [597]. Finally, Wallace served as Secretary of Commerce from 1945 to 1946.

[72] The best accounts of the origins and development of the Iowa State College Statistical Laboratory are *Statistics: An Appraisal*, edited by H.A. David and H.T. David [327], "Statistics in U.S. universities in 1933 and the establishment of the Statistical Laboratory at Iowa State" by H.A. David [324], "Highlights of some expansion years of the Iowa State Statistical Laboratory, 1947–72" by T.A. Bancroft [58], "Revisiting the past and anticipating the future" by O. Kempthorne [724], "The Iowa State Statistical Laboratory: Antecedents and early years" by H.A. David [322], and "Early statistics at Iowa State University" by J.L. Lush [859].

values are to be regarded as a sample is entirely avoided [1473, p. 165], and the ground-breaking work by Harold Hotelling and Margaret Pabst on rank data was designated to be completely distribution-free [653]. Bernard Welch, commenting on Fisher's *The Design of Experiments* in 1937, concluded that while the calculations required by exact inference would be lengthy, the result would be a test of hypothesis that was free of any assumptions [1428], and in 1938 Welch noted that an exact test of significance assumed nothing not obtained directly from the observed sample values [1429, p. 154].

E.J.G. Pitman, in his first of three papers, emphasized that the difference between two independent means could be determined without making any assumptions about the populations from which the samples were obtained; in the second paper on correlation, Pitman summarized the results of his investigation by stating that the test of significance made no assumptions about the sampled population; and in the third paper on analysis of variance, Pitman proposed a permutation test that involved no assumptions of normality, explaining that the observations were not to be regarded as a sample from a larger population [1129–1131]. Finally in 1938, Fisher in a little-known book published by the University of Calcutta Press, *Statistical Theory of Estimation*, was quoted as saying "it should be possible to draw valid conclusions from the data alone, and without a priori assumptions" [455, p. 23].

3. Associated with data-dependency and distribution-free alternatives to conventional tests, it was widely recognized that when utilizing permutation methods, samples need not be random samples from a specified population. Yates, discussing the Yeoman II wheat experiment, completely dismissed the notion that a sample of observations be drawn from a parent population [1473]. Also, Pitman noted in his discussion of the permutation version of the analysis of variance, that observations were not to be regarded as a sample from a larger population [1131]. Finally, Welch in his analysis of the correlation ratio, explained that he preferred to consider samples as drawn from a well-defined limited population rather than a hypothetical infinite population [1429].
4. It was generally accepted by many researchers that it was not necessary to calculate an entire statistic, such as a t or a z (later, F) when undertaking a permutation test. In fact, only that portion of the statistic that varied under permutation was required and the invariant portion could therefore be ignored, for permutation purposes. This recognition greatly reduced the computations necessary to perform an exact permutation test and allowed for more arrangements of the observed data to be considered in resampling permutation tests.

For example, Eden and Yates substantially reduced calculations by recognizing that the block and total sums of squares would be constant for all of their 1,000 samples and, consequently, the value of z for each sample would be uniquely defined by the treatment sum of squares, i.e., the treatment sum of squares was sufficient for a permutation analysis of variance test [379]. Welch, in his permutation analysis of randomized blocks, considered a monotonically increasing function of z that contained only the portion of z that varied under permutation. In this case, like Eden and Yates, Welch considered only the treatment sum of squares [1428]. Pitman, in his permutation analysis of two samples, observed that since the sample sizes (m and n) and grand mean ($\bar{z}$) were invariant over permutation of the observed data, each arrangement was a simple function of the sum of one sample for a one-sided probability value [1129].

Kendall and Babington Smith, in their discussion of the problem of m rankings, substantially reduced their calculations by recognizing that the number of rankings (m) and number of ranks (n) were invariant over permutation of the observed data and, therefore, calculated only the sum of squared deviations from the mean of the ranks in their permutation analysis of m rankings [739]. Likewise, Kendall, Kendall, and Babington Smith in their permutation analysis of Spearman's rank-order correlation coefficient, considered only the sum of the squared differences between ranks, which reduced computation considerably for each of the $n!$ arrangements of the observed rank-order statistics [746].

5. Yates developed a recursion process to generate hypergeometric probability values [1472] and Kendall utilized a recursion technique to generate successive frequency arrays of sums of concordant and discordant pairs for $n = 1, \ldots, 10$ [728]. Recursion methods were not new at this time, having been utilized historically by Blaise Pascal, Christiaan Huygens, James Bernoulli, Willem 'sGravesande, Pierre Rémond de Montmort, and Adolphe Quetelet, among others [571, 572]. Recursion methods were destined to become powerful tools for the production of exact probability values in the 1980s and 1990s when computers were finally able to generate complete discrete probability distributions with considerable speed and efficiency. It is important to mention recursion methods here as precursors to the algorithmic procedures employed by computer programmers in later decades.
6. Many of the permutation methods utilized by researchers in the 1920s and 1930s produced exact probability values based on all possible arrangements of the observed data values. For example, Fisher in his investigation of monozygotic and dizygotic twins calculated exact probability values based on all possible arrangements of Johannes Lange's data on twins and criminal activity [451]. Fisher also conducted an exact permutation analysis of the lady tasting tea experiment and an exact permutation analysis of Darwin's *Zea mays* data [451]. Hotelling and Pabst calculated exact probability values based on all $n!$ arrangements of the observed rank data, albeit for very small samples [653], and Friedman presented the exact distribution of χ_r^2 for a variety of values of p and n [485]. Pitman calculated exact probability values for his analysis of two-sample tests [1129]; Olds provided exact probability values for Spearman's rank-order correlation coefficient for values of $n = 2, \ldots, 7$ based on the $n!$ possible arrangements of one ranking against the other [1054]; Kendall constructed exact values of the differences between concordant and discordant pairs (Σ) for values of n from 1 to 10 [728]; and Kendall and Babington Smith created four tables of exact values for statistic W [739].

On the other hand, some researchers relied on a random sample of all possible arrangements of the observed data values, i.e., resampling-approximation probability values. While credit is usually given to Dwass in 1957 for the idea of resampling probability values [368], it is readily apparent that resampling was in use in the 1920s and 1930s, although in a rudimentary way. For example, Geary utilized a random sample of 1,000 arrangements of cell frequencies to establish the approximate probability of a correlation between potato consumption and the incidence of cancer [500], and Eden and Yates examined 1,000 out of a possible 4,586,471,424 arrangements of Yeoman II wheat shoots grown in eight blocks to generate an approximate probability value [379].

Something that was not emphasized in this chapter was the use of the method of moments to fit a continuous distribution to the discrete permutation distribution to obtain approximate probability values. The method of moments was typically used to generate probability values based on permutation distributions to compare with probability values obtained from asymptotic distributions, such as the normal or chi-squared distributions. For example, Pitman utilized a method of moments approach to obtain approximate probability values in all three of his seminal papers [1129–1131]. There, moments based on the observed data were equated to the moments of the beta distribution to obtain the correspondence between the probabilities of the observed statistic and probabilities from the associated beta distribution. Others who utilized moments of the permutation distribution to compare results to asymptotic distributions were Welch [1428] and Friedman [485] in 1937; Olds [1054] and Kendall [728] in 1938; and Kendall and Babington Smith [739], Kendall, Kendall, and Babington Smith [746], and McCarthy [911] in 1939.

7. Finally, the profusion of research on permutation methods for small samples by Hotelling and Pabst; Olds; Kendall and Babington Smith; and Kendall, Kendall, and Babington Smith ushered in the 1940s when tables of exact probability values were published for a number of statistics with small sample sizes. These early works constituted a harbinger of much of the work on permutation methods during the 1940s: a focus on creating tables for small samples that employed permutations for the calculations of exact probability values, primarily for rank tests.

1940–1959 3

The 1920s and 1930s constituted a time of early development for permutation statistical methods. This was also a period during which researchers recognized the difficulties of computing exact probability values for all but the smallest of data sets. Progress on the development of permutation methods continued over the next two decades, but in many ways that work took on a different focus from that of the previous two decades. The recognition of permutation methods as the gold standard against which conventional statistical methods were to be evaluated, while often implicit in the 1920s and 1930s, is manifest in many of the publications on permutation methods that appeared between 1940 and 1959. Also, a number of researchers turned their attention during this time period to rank tests, which simplified the calculation of exact probability values; other researchers continued work on calculating exact probability values, creating tables for small samples; and still others continued the theoretical work begun in the 1920s. What follows is first a brief overview of the achievements that took place in the two decades bridging the 1940s and 1950s, followed by an in-depth treatment of selected contributions. The chapter concludes with a look ahead at the rapid expansion of permutation statistical methods between 1960 and 1979.

3.1 Overview of This Chapter

The 1940s and 1950s saw a proliferation of non-parametric rank tests, which is not surprising since, strictly speaking, every rank test is a permutation test; although, not vice-versa; see for example, discussions by Feinstein in 1973 [421], Bradbury in 1987 [200], May and Hunter in 1993 [908], Good in 1994 and 2004 [523, 529], and Ernst in 2004 [413].[1] Examples of rank tests in this period include the Kendall rank-order correlation coefficient [728, 734]; the Friedman two-way analysis of

[1] A comprehensive overview of statistics in the 1950s is provided by Tertius de Wet in his presidential address to the South African Statistical Association in 2003 [335].

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9__3,
© Springer International Publishing Switzerland 2014

variance for ranks [485, 486], which is equivalent to the Kendall coefficient of concordance [734, 739, 1224, p. 335] and also to the Wallis correlation ratio for ranked data [1411]; the Wilcoxon two-sample rank-sum test [1453], independently developed by Mann and Whitney [880], Haldane and Smith [573], van der Reyden [1391], and Festinger, who, incidentally, was the first to accommodate unequal sample sizes [427]; the Wald–Wolfowitz runs test [1405]; the Jonckheere–Terpstra test for ordered alternatives [699, 1347]; the Mann test for trend [879]; the Kruskal–Wallis one-way analysis of variance rank test [779]; and the Mood median test [1001].

In addition, permutation methods were often employed to generate tables of exact probability values for small samples, e.g., tables for testing randomness by Swed and Eisenhart [1337]; for 2 × 2 contingency tables by Finney [434]; for the Spearman rank-order correlation coefficient by David, Kendall, and Stuart [328]; for the Wilcoxon two-sample rank-sum test by Wilcoxon [1453, 1454], White [1441], and Fix and Hodges [465]; for the Mann test for trend by Mann [879]; for a rank test of dispersion by Kamat [707]; and for the Mann–Whitney two-sample rank-sum test by van der Reyden [1391] and Auble [40].

A theme that was commonly repeated between 1940 and 1959 involved the difficulty of computing exact probability values for raw data and, in response, the conversion of the raw data to ranks to simplify computation. On this topic, in 1943 Scheffé [1230] introduced non-parametric randomization tests, building on the work of Fisher [448], remarking that "except for very small samples the calculation…[was] usually extremely tedious" [1230, p. 311], a problem that plagued permutation tests until the advent of high-speed computers. In that same year, Wald and Wolfowitz [1406] developed an exact test procedure for randomness based on serial correlation, which pointed the way for other researchers to develop derivations of asymptotic distributions for the non-rank case of the randomization method [1230, p. 311]. The Wald–Wolfowitz test provided an exact test of significance by enumerating all possible values of a test statistic for the measurement of serial correlation.

A year later, Wald and Wolfowitz [1407] devised exact tests of significance for use in cases when the form of the underlying probability distribution was unknown, extending the work done by R.A. Fisher in 1925 and 1935 [448, 451]. A general theorem on the limiting distribution of linear forms in the universe of permutations of observations was derived. Included in the discussion were applications to the Pitman test for two samples drawn from the same population [1129], the Pitman test for dependence between two variates [1130], the Welch [1428] and Pitman [1131] tests for randomized block designs, and Hotelling's T^2 generalization of Student's two-sample t test [652].

In 1948 Haldane and Smith provided an exact permutation test for birth-order defects, complete with tables [573]. This exact test was devised to test whether the probability of a child inheriting a certain medical condition, such as phenylketonuria, increased with birth order and was equivalent to the Wilcoxon two-sample rank-sum test. Pitman [1132] in unpublished, but widely circulated lecture notes for a course given at Columbia University in 1948, showed that the

Wilcoxon [1453] test for location had an asymptotic relative efficiency (ARE) of $3/\pi$ when compared to Student's t test under the assumption of normality. Also, Pitman showed that the Wald and Wolfowitz [1405] runs test had zero asymptotic efficiency for testing either location or dispersion [1001, p. 520]. New concepts introduced in these lectures included efficiency, asymptotic power, and asymptotic relative efficiency. Pitman's approach to ARE was first published by Noether in 1950 and extended by Noether in 1955 [1038, 1039]; see also a 2009 article on this topic by Lehmann [815]. Later, infinite classes of linear rank tests were introduced along with the distributions for which these tests were asymptotically most powerful for location and scale alternatives by Mielke in 1972 and 1974 [932, 933] and by Mielke and Sen in 1981 [987].

In 1949 Wolfowitz [1466] surveyed a number of problems in non-parametric inference and recommended that methods for obtaining critical regions be developed in connection with the randomization methods of Fisher [448] and Pitman [1129]. Lehmann and Stein showed that the permutation tests introduced by Pitman [1129–1131], when applied to certain discrete problems, coincided with the Fisher two-sample permutation test and that the two-sample permutation test of Pitman [1129] was most powerful against the alternative that the two samples were independently normally-distributed with common variance [818]. In 1951 Freeman and Halton [480], in what would later become a landmark article, described an exact test for small samples in $r \times c$ and $2 \times 2 \times 2$ contingency tables when the chi-squared test of independence was not applicable.

In 1952 Wassily Hoeffding (also, Höffding) investigated the power of a family of non-parametric tests based on permutations of observations, finding the permutation tests to be asymptotically as powerful as the related parametric tests [636]. These tests included the Pitman tests for two independent samples [1129], bivariate correlation [1130], and randomized blocks analysis of variance [1131]; the Fisher analysis of variance [451]; and the Welch test for randomized blocks [1428]. This was a recurring theme that was also addressed by Silvey in 1953 and 1954, who further considered the problem of determining the conditions under which the permutation distribution of a statistic and its normal-theory distribution were asymptotically equivalent [1275, 1276].[2] As detailed by Baker and Collier [52], Silvey showed analytically that the permutation distribution of the Fisher variance-ratio z statistic for a one-factor treatment arrangement was asymptotically the F distribution [52]. In 1955 Box and Andersen also discussed the use of permutation tests to assess the effect of departures from normality on standard statistical tests, with specific references to the one-way analysis of variance and randomized block designs [193]. Similarly, see a 1973 article by Robinson who considered the same problem as Hoeffding, but did not assume that the errors were independently distributed with equal variances [1178].

[2] Unfortunately, these two important articles by Samuel Silvey went largely unnoticed, published as they were in *Proceedings of the Glasgow Mathematical Association*, a journal that was not widely distributed at the time.

In 1955 Hack generated an empirical F-ratio distribution based on 100 random selections of the possible permutations of 80 values for each of two root depths for tomato plants grown under greenhouse conditions [566]. Hack found that when the data were approximately normally-distributed, Snedecor's F-ratio and the permutation version of the analysis of variance F test generally agreed, but when the data were skewed, a deficiency of large and small values of F underestimated significance at the 5 % level. Hack included in his study one data set with skewness coefficient $g_1 = 1.5$, and kurtosis coefficient $g_2 = 3.0$, and a second set of data with $g_1 = 3.6$ and $g_2 = 15.9$. As documented by Baker and Collier [52], only for the latter set did the empirical permutation distributions of the variance ratios differ noticeably from the corresponding F distributions under normal-theory methods [52]. In 1958 Johnson [690] found the empirical distribution of the F-ratio to be similar to the randomization distribution studied by Welch in 1937 [1428].

Also in 1955, Kempthorne described the use of randomization in experimental designs and how randomization permitted evaluation of the experimental results [719]. Included in his discussion were analysis of variance procedures for completely randomized, randomized block, and Latin square designs. In 1956 Kamat [707] proposed a test for the equivalence of two parameters of dispersion, based on ranks, which was a modification of the Mann–Whitney two-sample rank-sum test [880]. Kamat also included tables for selected significance levels for small samples.

In 1956 Scheffé discussed alternative permutation models for the analysis of variance [1231]. Under the heading of "randomization models," Scheffé provided an insightful comparison of the ordinary analysis of variance and the permutation version of the analysis of variance. The following year, 1957, Dwass [368] continued the general theme of computational difficulties for permutation tests, even with small samples. In the same manner as Eden and Yates in 1933 [379], Dwass recommended taking random samples of all possible permutations for a two-sample test and making the decision to reject or fail to reject the null hypothesis on the basis of these random permutations only.

In 1958 Sawrey published a short paper on the distinction between exact and approximate non-parametric methods, with the first leading to an exact significance level and the second to an approximate significance level [1227]. Sawrey cautioned future researchers on the importance of the differences and concluded that when an exact permutation test was available, it should "always be used unless the labor is completely prohibitive" [1227, p. 175].

Also in 1958 Chung and Fraser proposed a number of randomization tests for multivariate two-sample problems [254]. Noting that with few observations on a large number of variables the Hotelling generalized T^2 test cannot be computed, they proposed several alternative tests based on permutation methods. Finally, like Dwass [368], they observed that valid permutation tests could be obtained from a subgroup of all possible permutations, thereby substantially reducing the amount of computation required.

3.2 Development of Computing

Because permutation tests are inherently computationally-intensive, it took the development of high-speed computing for permutation tests to achieve their potential. What few computers were available in the period between 1940 and 1959 were large, slow, very expensive to use, and located at only a few computing centers. Moreover, in large part their use was restricted to military and industrial applications and thus were not generally accessible to those involved in the development of permutation methods. John Vincent Atanasoff at Iowa State University, with the assistance of his graduate student Clifford Berry, is usually credited with inventing the first automatic electronic digital computer, which was fully completed in 1942. Atanasoff, in an attempt to justify the construction of a computer, described the problems it could be expected to solve.[3] Although Atanasoff was an engineer and applied scientist, the first three problems on his grant request to Iowa State College were statistical problems: multiple correlation, curve fitting, and the method of least squares. Other problems on his list included questions relating to quantum mechanics, electric circuit analysis, elasticity, and other problems primarily of interest to engineers.

By the late 1930s punched-card machine technology had become so well established and reliable that Howard Aiken, a graduate student in theoretical physics at Harvard University, in collaboration with engineers at IBM, undertook construction of a large automatic digital computer that eventually became known as the Harvard Mark I.[4] The Mark I was the largest electro-mechanical calculator ever built. It was a behemoth of a machine that was 51 ft long, 3 ft deep, 8 ft high, weighed nearly five tons, possessed 765,000 components, and contained 530 miles of wiring. The Mark I was completed in 1944, but its use was largely restricted to producing mathematical tables. The Mark I was superseded by the Mark II in 1948, the Mark III in 1949, and the Mark IV in 1952.

J. Cornfield and the Mark I

Salsburg relates an interesting anecdote about Jerome Cornfield and the Mark I which illustrates both the expense and limited access of computers at that time [1218]; see also a chapter by Hilbe for another version of this story [618]. In the late 1940s, Jerome Cornfield at the Bureau of Labor Statistics had a mathematical problem. Cornfield needed to invert a 24×24 matrix for Wassily

(continued)

[3] Atanasoff's proposal for construction of the computer was funded by Iowa State College, (now, Iowa State University) which granted Atanasoff $5,000 to complete his computing machine.

[4] Technically, the computer was originally called the Aiken–IBM Automatic Sequence Controlled Calculator (ASCC) and was renamed the Mark I by Harvard University when it was acquired from IBM on 7 August 1944.

Leontief, the Nobel prize-winning Columbia University economist he was working with at the time, but the estimated time for him to invert the matrix by hand was 100 years working 12 h a day. Cornfield and Leontief decided to send their 24 × 24 matrix to Harvard University to have it inverted on the Mark I. When they contracted to pay for the project, the funding was denied by the Bureau of Labor Statistics on the grounds that the government would pay for "goods," but not for "services." Cornfield then negotiated with the Bureau for a purchase order for capital goods. The invoice called for "one matrix, inverted" [1218, pp. 177–179]. The matrix was successfully inverted by the Mark I, taking only several days instead of 100 years [618, pp. 5–6].

The IAS (Institute for Advanced Study) computer was built from late 1945 to 1951 under the direction of John von Neumann (originally, Neumann János Lajos) for the Institute for Advanced Study in Princeton, New Jersey. The IAS computer was a stored-program parallel-processor computer and the architectural design was so successful that most computers built since the 1940s have been "von Neumann" machines [41, Sect. 5.8]. The IAS was a binary computer with a 40-bit word and 1,024 words of memory. It could perform 2,000 multiplications in one second and add or subtract 100,000 times in the same period [240, p. 278]. In March of 1953 there were only 53 kilobytes of high-speed random-access memory in the entire world; five kilobytes (40,960 bits) were housed in the IAS computer [370, p. 4]. When President Eisenhower appointed von Neumann to the Atomic Energy Commission (AEC) in 1954, von Neumann left the Institute and the computer project went into decline. Three years later, on 8 February 1957, John von Neumann died of advanced metastasizing cancer; he was only 53 years of age.[5] As George Dyson reported, without its messiah, the computer project at the Institute of Advanced Study lost support and was terminated. At midnight on 15 July 1958, Julian Himely Bigelow, von Neumann's chief engineer, turned off the master control, logged off with his initials, J.H.B, and The Institute for Advanced Study Numerical Computing Machine ceased functioning [370, p. 315].

The year 1946 saw the completion of ENIAC (Electronic Numerical Integrator and Computer), the first general purpose computer built for the United States Army Ballistic Research Laboratory in the Moore School of Engineering at the University of Pennsylvania with a speed of 5,000 simple additions or subtractions per second. The ENIAC computer contained 17,648 double-triode vacuum tubes, had 500,000 soldered joints, 1,500 hundred relays, hundreds of thousands of resistors, capacitors, and inductors, weighed 27 tons, and occupied 680 ft^2 of space [1424]. In 1949 Andrew Hamilton famously predicted that "[w]here a calculator like the ENIAC today is equipped with 18,000 vacuum tubes and weighs 30 tons, computers in the

[5] For an interesting biography of John von Neumann, as related to computers and computing science, see Chap. 4 in the 2012 book *Turing's Cathedral* by George Dyson [370, Chap. 4].

future may have only 1,000 vacuum tubes and perhaps weigh only $1\frac{1}{2}$ tons" [580, p. 258]. In late 1947 the ENIAC was moved 200 miles to its permanent home at the Ballistics Research Laboratory at Aberdeen Proving Ground in Maryland.

In 1947, the transistor was invented by William Shockley, Walter Brattain, and John Bardeen at Bell Laboratories (now, Alcatel–Lucent) in Murray Hill, New Jersey.[6] However, it was not until 1956 that the first transistorized computer was constructed; it was named the TX-0 by its designers at the Massachusetts Institute of Technology [618]. The impact of the transistor on computing cannot be overstated. When Bell Laboratories announced the invention of the transistor in 1948, the press release boasted that more than a hundred transistors could easily be held in the palm of the hand. Today, a person can hold more than 100 billion transistors in the palm of one hand. Moreover, on today's market, transistors cost only about a dollar per billion, making them the cheapest and most abundant manufactured commodity in human history [602, p. 106].

In 1949 the EDSAC (Electronic Delay Storage Automatic Calculator) computer successfully ran its first program at the University of Cambridge, computing all the squares of numbers from 0 to 99. EDSAC was a general purpose serial electronic calculating machine installed at the Cambridge University Mathematical Laboratory. EDSAC could process 650 instructions per second with 1,024 17-bit words of memory stored in mercury delay lines, each of which was about 5 ft long, and ran at 500 kHz with a multiplication time of about 7 ms. In 1950 EDSAC began providing general service to the University of Cambridge users.

Fisher and Computing

Many people have surmised what R.A. Fisher could have accomplished if only he had access to a modern computer. As noted by Edwards in 2012 [401, p. 44], in 1950 Fisher was the first person to tackle a biological problem with a computer, publishing the results in an article titled "Gene frequencies in a cline determined by selection and diffusion" in *Biometrics* [458]. The analysis required the tabulation of the solution of a second-order non-linear differential equation with two boundary point conditions; consequently, he called on his friend Maurice Wilkes, the constructor of the EDSAC computer at the University of Cambridge, who passed the problem to one of his students, David Wheeler, in whose Ph.D. thesis the solution first appears [401, p. 44]. In Fisher's own words:

> [v]alues of q [one minus the probability] to eight decimal places, from $x = 0$, by intervals of .02, to extinction, are given in Table I. I owe this tabulation to

(continued)

[6] This statement paints a rosy picture of the relationship between Shockley, on the one hand, and Brattain and Bardeen, on the other hand, that was nothing but congenial. For a more detailed account, see a 2010 book by Sam Kean titled *The Disappearing Spoon* [712, pp. 41–43].

Dr. M. V. Wilkes and Mr. D. J. Wheeler, operating the EDSAC electronic computer. The last decimal place may be in error by 3 or 4 digits [458, p. 357].

It was a difficult solution as it involved programming of an automatic trial-and-error method for satisfying the boundary conditions at the two ends of the interval. The entire story is related in the autobiography of Maurice V. Wilkes, *Memoirs of a Computer Pioneer*, published in 1985 [1455, pp. 148–149].

It is interesting to compare the advances in computing in the United States after World War II with those of Great Britain. The two models, one based on a cooperative effort between private industry and the federal government, and the other based on the federal government alone, provide a vivid contrast in the speed with which computing was adopted by both universities and private corporations in both countries.

Computing in Great Britain

When World War II broke out, British mathematicians and physicists were enlisted to work on the development of intelligence and early warning systems in government laboratories and institutions. Two of the institutions were the Telecommunications Research Establishment in Great Malvern and the highly secretive Government Code and Cipher School at Bletchley Park [426, p. 53]. As Georgina Ferry related in a book titled *A Computer Called LEO*, at that time Great Britain was actually ahead of the United States in developing computing capability. The Mark I Colossus, the first truly electronic programmable digital computer, was developed by Thomas H. (Tommy) Flowers at Bletchley Park in 1943, and was followed six months later by the Mark II Colossus, which was five times faster than the Mark I Colossus. Thus, electronic computers had been built and were working in Great Britain while ENIAC was still on the drawing boards at the Moore School of Engineering at the University of Pennsylvania [426, pp. 55–57].

The development and construction of computers in the United States was often a joint effort between the government and the private sector, including Bell Laboratories and such universities as the University of Pennsylvania, Iowa State University, and Harvard University. Thus, while ENIAC was quickly declassified and achieved world fame in the post-war years, Colossus was a military project and remained obscured behind the impervious wall of the Official Secrets Act [426, p. 56]. As Newton E. Morton described in an obituary of Cedric Smith, human genetics and other scientific fields in Great Britain were "[expletive deleted] by government policy that protested British computers [and] for a score of years the sciences that needed competitive

(continued)

computing were stifled, and many of their practitioners changed disciplines or countries" [1008, p. 10]. Thus, there was no concerted effort by the British government to explore and develop civilian applications [426, p. 57]. It was not until 30 years later in 1974 that the secrecy act was lifted.[7] In the meantime, Tommy Flowers had done as he was told and burned all of his records. Today, the Bletchley Park Trust proudly shows visitors around the site and in 1996 a reconstruction of Colossus was unveiled [426, p. 57].

In 1951, the first UNIVAC (UNIVersal Automatic Computer) computer was delivered to the United States Census Bureau, with a speed of 1,905 operations per second. In 1952 Univac Computer Corporation applied for and eventually received a patent on the digital computer, which was voided in the late 1960s when Honeywell Computer Corporation sued Univac claiming that Univac did not have a right to a patent on computers [556]. It was a UNIVAC 1 computer at the United States Census Bureau that provided, for the first time, a computer-based forecast of the 1952 U.S. Presidential election between Dwight D. Eisenhower and Adlai E. Stevenson. It also was the first time that a working computer was shown on television, as the returns were broadcast by the Columbia Broadcasting System (CBS) in November of 1952 [556].

A year earlier, in November of 1951, LEO (Lyons Electronic Office) became the first computer in the world to be harnessed to the task of managing a business, anticipating IBM by 5 years. That business was J. Lyons & Company, renowned throughout England for its fine teas and cakes [426, p. viii]. In 1952, the MANIAC (Mathematical Analyzer, Numerical Integrator, And Computer) computer was installed at the Los Alamos Scientific Laboratory in New Mexico. MANIAC was an all vacuum-tube computer primarily used for "Project Mike" in the development of the first hydrogen bomb. MANIAC had 1,024 words of memory with a word length of 40 bits and, in addition, had a 10,000 word drum for auxiliary storage [1436]. Maniac was later upgraded to five kilobytes of memory. As George Dyson observed in 2012, that is less than what is allocated to displaying a single icon on a computer screen today [370, p. ix].

In 1953 the ORACLE (Oak Ridge Automatic Computer and Logical Engine) computer was installed at the Clinton Engineer Works in Oak Ridge, Tennessee. ORACLE was based on the IAS architecture developed by John von Neumann and used both vacuum tubes and transistors (q.v. page 106). ORACLE employed a Williams tube for 1,024 words of memory of 40 bits each (later doubled to 2,048 words) and, at the time, was the fastest computer and possessed the largest data

[7] In fact, it took 70 years for the Government Communications Headquarters (GCHQ) to release two papers written by Alan Turing between April 1941 and April 1942 while he was head of wartime code-breaking at Bletchley Park. The two papers on "Paper on the statistics of repetitions" and "Other applications of probability to cryptography" were finally released in April of 2012 [25].

storage capacity of any computer in the world [753]. In 1953, the first magnetic core memory was installed at the Massachusetts Institute of Technology and IBM shipped its first computer, the IBM 701, with a speed of 16,000 operations per second. In 1957, FORTRAN (FORmula TRANslation) was developed by John Warner Backus at IBM,[8] and in 1958 Jack Kilby created the first monolithic integrated circuit at Texas Instruments in Dallas, Texas. At the same time, Robert Noyce at Fairchild Semiconductor in Mountain View, California, independently created the integrated circuit. Kilby made his integrated circuit with a germanium (Ge) surface, while Noyce made his with a surface of silicon oxide (SiO_2) [618]. The first computer hard drive was developed by IBM in 1956; it consisted of 50 two-foot diameter platters, could store five million characters, and weighed one ton [618].

It is, perhaps, interesting to note that in the mid-1950s computers contained either built-in pseudorandom number generators (PRNG) or could refer to random number tables [1419]. Unfortunately, the initial random number (seed) in the standard FORTRAN library was a constant, meaning that all simulations using this subroutine were using the same series of random numbers [1344, p. 43].

No account detailing the development of computing in this period would be complete without a mention of Rear Admiral Grace Hopper, programmer of, at that time, the world's most complex computing machine, the "mother of COBOL," and the first woman to earn a Ph.D. in mathematics from Yale University in the school's 223 year history [165, pp. 25–26].[9]

Grace Hopper

Grace Brewster Murray Hopper graduated Phi Beta Kappa with a degree in mathematics and physics from Vassar College in Poughkeepsie, New York, in 1928, then earned her M.A. in mathematics from Yale University in 1930 and her Ph.D. in mathematics from Yale University in 1934 under the direction of algebraist Øystein Ore. In 1934 Hopper accepted a full-time academic position at her undergraduate alma mater, Vassar. In 1940 Hopper took a 1-year sabbatical to study with the celebrated mathematician Richard Courant at New York University. In the fall of 1942 Hopper returned to her tenured position at Vassar. However, in late 1943 Hopper took a leave of absence from her position at Vassar, making a life-altering decision to serve her country by joining the U.S. Navy. She reported to the United States Naval Reserve

(continued)

[8] For a history of the development of FORTRAN, see the recollection by John Backus in the special issue of *ACM SIGPLAN Notices* on the history of programming [44].

[9] This statement is from Kurt Beyer, *Grace Hopper and the Invention of the Information Age*. Actually, the first woman to earn a Ph.D. in mathematics from Yale University was Charlotte Cynthia Barnum (1860–1934) who received her Ph.D. in mathematics in 1895 [1018].

Midshipmen's School–Women in Northampton, Massachusetts, in December of 1943. Hopper graduated from Midshipmen's School in 1944 as battalion commander and first in her class [165].

Much to her surprise, upon graduation Lieutenant (Junior Grade) Hopper was assigned to the Bureau of Ordnance Computation at Harvard University, becoming the third programmer of the world's most complex, unique computing machine, the Automatic Sequence Controlled Calculator (ASCC), later renamed the Harvard Mark I. The ASCC was an early example of a programmable machine and was housed, under high security, in the basement of Harvard University's Cruft Physics Laboratory. Hopper worked under Commander Howard H. Aiken and it was here that she was credited with coining the term "bug" in reference to a glitch in the computer: actually a large moth had flown into the laboratory through an open window on 9 September 1945 and was stuck between points at Relay #70, Panel F, of the Mark II Aiken Relay Computer, whereupon she remarked that they were "debugging" the system [870, 1042].

At the conclusion of the war, Hopper resigned from Vassar to become a research fellow in engineering and applied physics at Harvard's Computation Laboratory and in 1949 she joined the Eckert–Mauchly Computer Corporation as a senior mathematician, retaining her Naval Reserve commission. The corporation was soon purchased by Remington Rand in 1950, which merged into the Sperry Corporation in 1955. Here Hopper designed the first compiler, A-0, which translated symbolic mathematical code into machine code [870]. In 1966, then Commander Hopper retired from the Naval Reserves, but was recalled less than seven months later. In 1973 Hopper was promoted to the rank of Captain, in 1983 she was promoted to the rank of Commodore in a ceremony at the White House, and in 1985 she was elevated to the rank of Rear Admiral. In 1986, after 43 years of military service, Rear Admiral Grace Hopper retired from the U.S. Navy on the deck of the USS Constitution at the age of 80. She spent the remainder of her life as a senior consultant to the Digital Equipment Corporation (DEC). Grace Brewster Murray Hopper died in her sleep on 1 January 1992 at the age of 86 and was buried with full military honors at Arlington National Cemetery in Arlington, Virginia [823, 1042].

3.3 Kendall–Babington Smith and Paired Comparisons

In 1940 Maurice G. Kendall (q.v. page 84) and Bernard Babington Smith, the renowned University of Oxford experimental psychologist, published a lengthy article on the method of paired comparisons [741]. This article dealt with the same problem as their 1939 article (q.v. page 86) on the problem of m rankings [739], but in a very different manner. They considered a general method of investigating preferences. As they explained, given n objects suppose that each of the $\binom{n}{2}$ possible pairs is presented to an observer and the preference of one member of each pair is recorded. With m observers the data then comprise $m\binom{n}{2}$ preferences. The primary question for Kendall and Babington Smith was: is there any significant concordance of preferences between observers? [741, p. 325].

B. Babington Smith

Bernard Babington Smith, known as BBS, was one of four sons and five daughters of Sir Henry Babington Smith and Lady Elizabeth Babington Smith (née Bruce), daughter of the 9th Earl of Elgin (Victor Alexander Bruce), grandson of the Lord Elgin (Thomas Bruce) who removed the Elgin marbles from the Parthenon and other buildings on the Acropolis of Athens while he was serving as the British ambassador to the Ottoman Empire from 1799 to 1803. Little is known of the early years of Bernard Babington Smith, but he was most likely home-schooled like his sister, Constance Babington Smith. In 1939 Babington Smith was a Lecturer in Experimental Psychology at the University of St. Andrews in Scotland, but resigned to join the Royal Air Force (RAF) at the beginning of World War II. He served with his celebrated sister, Constance Babington Smith, and with Sarah Oliver, daughter of Sir Winston Churchill, as wartime photographic interpreters in the Allied Photographic Intelligence Unit at Medmenham, Buckinghamshire. It was Constance Babington Smith who first identified a pilotless aircraft at Peenemünde, a major German rocket research facility on the Baltic Coast, and it was her discovery that led to a critical bombing campaign by Allied Forces that flattened strategic launch sites in France. Constance Babington Smith writes about her brother in her book *Evidence in Camera*, and supplies a picture of Bernard Babington Smith [42].

After the war Babington Smith joined the faculty at the University of Oxford in 1946. While at Oxford, Babington Smith collaborated with Maurice Kendall (q.v. page 84) on a number of projects that resulted in important publications on ranking methods and random numbers. One of his students was Ralph Coverdale, founder of the Coverdale Organisation. Coverdale and Babington Smith worked together for many years on Coverdale Training, a highly developed form of learning through action. In 1973 when Babington Smith retired, he was Senior Lecturer in Experimental Psychology and Fellow of Pembroke College at the University of Oxford. Bernard Babington Smith died on 24 August 1993 at the age of 88 [683, 808].

Consider a set of n objects $\{\omega_1, \ldots, \omega_n\}$ and an observer who is asked to choose between every pair. If ω_1 is preferred over ω_2, write $\omega_1 \to \omega_2$. If the observer is not completely consistent, preferences might be made of the type $\omega_1 \to \omega_2 \to \omega_3 \to \omega_1$. This they termed an inconsistent or circular triad. Let d be the number of circular triads in a given experiment, then Kendall and Babington Smith showed that

$$\zeta = 1 - \frac{24d}{n^3 - n} \qquad \text{if } n \text{ is odd}$$

and

$$\zeta = 1 - \frac{24d}{n^3 - 4n} \qquad \text{if } n \text{ is even}$$

may be regarded as a coefficient of consistency, with $0 \leq \zeta \leq 1$ and $n \geq 3$.

For a brief aside, consider that 10 years later in 1950, B. Babington Smith served as a discussant at a symposium on ranking methods organized by the Royal Statistical Society with presenters P.A.P. Moran, J.W. Whitfield, and H.E. Daniels [314, 1005, 1444]. Here Babington Smith expressed caution regarding the use of paired ranking methods, in general, and the coefficient of consistency, in particular. He related watching a subject rank order nine items and observed that the subject made more than 70 comparisons between pairs of items.[10,11] As to the coefficient of consistency, Babington Smith suggested a new definition and a new symbol for the coefficient, where

$$\Xi = 1 - \binom{n}{2}^{-1} 4d \ ,$$

which gives the same form, but a different minimum value, for n odd and even. The advantage, he noted, is that Ξ is more in line with other coefficients and when the chance expectation of circular triads is realized, the value of Ξ is zero. Finally, he observed that there is a certain advantage to attaching a negative sign to the situation where the number of circular triads exceeds chance expectation.

Based on the permutation structure of ζ, Kendall and Babington Smith calculated the exact probability distribution of ζ for $n = 3, \ldots, 7$, and conjectured that the four moments of ζ were given by

[10] The maximum number of inversions required to reverse the order of ranks follows an irregular series. For $n = 2$, the maximum number of inversions, I, is 1; for $n = 3$, $I = 3$; for $n = 4$, $I = 6$; and so on. Thus, the sequence is 1, 3, 6, 10, 15, 21, and so on. The sequence is a component of Pascal's triangle; see Column 3 in Table 3.11 of this chapter, page 186. Any successive number can be obtained by $n(n + 1)/2$. Thus, for $n = 9$ objects the maximum number of inversions is $9(9+1)/2 = 45$. It stretches the imagination that a subject made more than 70 paired comparisons to rank order only nine objects.

[11] This is the reason that the Academy of Motion Picture Arts and Sciences places a maximum limit of 10 nominations for the Academy Awards (Oscars), as it is too difficult for the judges to rank order a larger number of nominations.

$$\mu_1' = \frac{1}{4}\binom{n}{3}\,,$$

$$\mu_2 = \frac{3}{16}\binom{n}{3}\,,$$

$$\mu_3 = -\frac{3}{32}\binom{n}{3}(n-4)\,,$$

and

$$\mu_4 = \frac{3}{256}\binom{n}{3}\left[9\binom{n-3}{3} + 39\binom{n-3}{2} + 9\binom{n-3}{1} + 7\right]\,,$$

these being polynomials in n which agreed with their numerical calculations for $n = 3, \ldots, 7$. If $m < s$, then $\binom{m}{s} = 0$. Kendall and Babington Smith explained that they had very little doubt that the moments were correct, but were unable to offer a rigorous proof [741, p. 332]. They further conjectured that the distribution of ζ tended to normality as n increased.[12]

3.4 Dixon and a Two-Sample Rank Test

In 1940 Wilfrid Dixon devised a statistic that he called C^2. The new statistic was designed to test the null hypothesis that two samples represent populations with the same distribution function [353].

W.J. Dixon

Wilfrid Joseph Dixon received his B.A. degree in mathematics from Oregon State College (now, Oregon State University) in 1938, his M.A. degree in mathematics from the University of Wisconsin in 1939, and his Ph.D. in mathematical statistics from Princeton University in 1944 under the supervision of Samuel S. Wilks. Dixon accepted a position at the University of Oklahoma in 1942, moved to the University of Oregon in 1946, and moved again to the University of California at Los Angeles (UCLA) in 1955, where he remained until his retirement in 1986. While at UCLA, Dixon formed the Biostatistics

(continued)

[12] These results were later proved by Patrick Moran in a brief article on "The method of paired comparisons" in *Biometrika* in 1947 [1003].

Division in the School of Public Health and also organized and chaired the Department of Biomathematics in the School of Medicine. Wilfrid Joseph Dixon died at home from heart disease on 20 September 2008 at the age of 92 [472, 473].

In 1940 Dixon published a short note in *The Annals of Mathematical Statistics* on a criterion for testing the null hypothesis that two samples have been drawn from populations with the same distribution function [353]. Following the notation of Dixon [353], let the two samples, O_n and O_m, be of size n and m, respectively and assume $n \leq m$. Arrange in order the elements $u_1, \ldots, u_n$ of O_n into their order statistics, i.e., $u_1 < u_2 < \cdots < u_n$, where the elements represent points along a line. The elements of the second sample, O_m, represented as points on the same line are then divided into $n + 1$ groupings by the first sample, O_n in the following manner: let m_1 be the number of points with a value u_1, m_i is the number of points with a value greater than u_i and less than or equal to u_{i+1} for $i = 1, \ldots, n$, and m_{n+1} is the number of points with a value greater than u_n. The criterion proposed by Dixon was

$$C^2 = \sum_{i=1}^{n+1} \left(\frac{1}{n+1} - \frac{m_i}{m} \right)^2 . \tag{3.1}$$

An example will illustrate the calculation of C^2. Consider samples $O_n = \{3, 9, 1, 5\}$ with $n = 4$ elements, $O_m = \{6, 2, 8, 7, 2\}$ with $m = 5$ elements, and arrange the elements in order representing points along a line:

$$\underline{1},\ \ 2,\ \ 2,\ \ \underline{3},\ \ \underline{5},\ \ 6,\ \ 7,\ \ 8,\ \ \underline{9},$$

where an underline indicates the element is from the first sample, O_n. Then $m_1 = 0$, as there are no points less than $u_1 = 1$; $m_2 = 2$, as there are two points (2 and 2) between $u_1 = 1$ and $u_2 = 3$; $m_3 = 0$, as there are no points between $u_2 = 3$ and $u_3 = 5$; $m_4 = 3$, as there are three points (6, 7, and 8) between $u_3 = 5$ and $u_4 = 9$; and $m_5 = 0$, as there are no points greater than $u_4 = 9$. Then, following Eq. (3.1),

$$\begin{aligned} C^2 &= 3\left(\frac{1}{4+1} - \frac{0}{5} \right)^2 + \left(\frac{1}{4+1} - \frac{2}{5} \right)^2 + \left(\frac{1}{4+1} - \frac{3}{5} \right)^2 \\ &= 3(0.02)^2 + (-0.20)^2 + (-0.40)^2 \\ &= 0.60 + 0.04 + 0.16 \\ &= 0.80 \ . \end{aligned}$$

Dixon provided a table for values of C^2 for $m \leq 10$, $n \leq 10$, and $\alpha = 0.01, 0.05$, and 0.10 where C_α^2 was defined as the smallest value of C^2 for which $P(C^2 \geq C_\alpha^2) \leq \alpha$.[13]

Dixon showed that if m and n increased indefinitely in the ratio $n/m = \gamma$, then nC^2 converged stochastically to $\gamma + 1$, whereas if n is small, nC^2 ranged from 0 to $n^2/(n+1)$, indicating a distribution with a tail to the right. To illustrate the range of nC^2, consider two samples $O_n = \{1, 2, 3, 4\}$ with $n = 4$ elements, $O_m = \{5, 6, 7, 8, 9\}$ with $m = 5$ elements, arranged in order representing points along a line:

$$\underline{1},\ \ \underline{2},\ \ \underline{3},\ \ \underline{4},\ \ 5,\ \ 6,\ \ 7,\ \ 8,\ \ 9,$$

where an underline indicates the element is from the first sample, O_n. Then, $m_1 = m_2 = m_3 = m_4 = 0$, $m_5 = 5$, and following Eq. (3.1),

$$\begin{aligned} C^2 &= 4\left(\frac{1}{4+1} - \frac{0}{5}\right)^2 + \left(\frac{1}{4+1} - \frac{5}{5}\right)^2 \\ &= 4(0.20)^2 + (-0.80)^2 \\ &= 0.16 + 0.64 \\ &= 0.80\ . \end{aligned}$$

Then, $nC^2 = 4(0.80) = 3.20$, which is equal to $n^2/(n+1) = 4^2/(4+1) = 3.20$, the upper limit of the range given by Dixon [353, p. 202].

For the lower limit of the range, consider two samples $O_n = \{2, 4, 6, 8\}$ with $n = 4$ elements consisting of four consecutive even integers, $O_m = \{1, 3, 5, 7, 9\}$ with $m = 5$ elements consisting of five consecutive odd integers, arranged in order representing points along a line:

$$1,\ \ \underline{2},\ \ 3,\ \ \underline{4},\ \ 5,\ \ \underline{6},\ \ 7,\ \ \underline{8},\ \ 9,$$

where an underline indicates the element is from the first sample, O_n. Then $m_1 = m_2 = m_3 = m_4 = m_5 = 1$, and following Eq. (3.1),

$$C^2 = 5\left(\frac{1}{4+1} - \frac{1}{5}\right)^2 = 5(0)^2 = 0.00\ .$$

Thus, when the two samples are in random order with respect to each other, the lower limit is $nC^2 = 4(0.00) = 0$. This suggested to Dixon that for larger samples of m and n it was reasonable to try to fit the distribution of nC^2 by the method of

[13] This was a typical approach for the time. Because a permutation test generally did not generate a value of the statistic that coincided exactly with α (e.g., 0.05 or 0.01) of the permutation distribution, a value of the permuted statistic, C_α^2, was defined as the smallest value of statistic C^2 for which $P(C^2 \geq C_\alpha^2) \leq \alpha$.

moments. The rest of this short article by Dixon was devoted to finding the moments of nC^2 and fitting a continuous probability distribution, in this case the chi-squared distribution with v degrees of freedom, where

$$v = \frac{an(n+m+1)}{n+1}$$

and

$$a = \frac{m(n+3)(n+4)}{2(m-1)(m+n+2)(n+1)} \; .$$

3.5 Swed–Eisenhart and Tables for the Runs Test

In 1943 Frieda Swed and Churchill Eisenhart, both of whom were at the University of Wisconsin at the time, developed a runs test that was based on the existing runs test of Wald and Wolfowitz [1337].

F.S. Swed

Little is known of Frieda Selma Swed after she earned her B.A. and M.A. in mathematics from the University of Wisconsin at Madison in 1935 and 1936, respectively. In 1937 she was appointed as a research assistant in Agricultural Economics and in 1942 as a research assistant in the Agricultural Statistical Service at the University of Wisconsin. On 17 March 1946 she married Herbert E. Cohn who was an accountant for the University of Wisconsin. Herbert Cohn passed away on 27 June 1995 at age 81. Frieda Cohn worked for the University of Wisconsin for 50 years, principally for the Numerical Analysis Laboratory. She was an ardent University of Wisconsin booster, who for many years tutored the University of Wisconsin athletes in calculus and higher mathematics. Frieda Swed Cohn passed away on 7 October 2003 at the age of 88 [332, p. 27].

C. Eisenhart

Churchill Eisenhart received his A.B. degree in mathematical physics in 1934 and a year later, his A.M. degree in mathematics, earning both from Princeton University. During his junior year, Eisenhart was introduced to statistics when his professor, Dr. Robert U. Condon, gave him a copy of *Statistical Methods*

(continued)

for Research Workers by R.A. Fisher. As a physics major, the Fisher text made Eisenhart realize that "most physicists simply do not know how to handle small sets of measurements" [1057, p. 513] and the book kindled his interest in statistics. Eisenhart worked with Samuel Wilks while he was at Princeton, and it was Wilks who suggested that Eisenhart move to University College, London, where Egon Pearson was chair, for his Ph.D. studies.

Eisenhart completed his Ph.D. at University College, London, in 1937 under the direction of Jerzy Neyman. Upon graduation, Eisenhart returned to the United States, taking a position at the University of Wisconsin at Madison, where he remained until 1947, although during World War II Eisenhart was a Research Associate at Tufts University, a Research Mathematician in the Applied Mathematics Group at Columbia University, and Principal Mathematical Statistician for the Statistical Research Group at Columbia (q.v. page 69). In 1945, Condon was appointed head of the National Bureau of Standards (NBS)[14] and brought Eisenhart to the Bureau in October of 1946. Eisenhart was appointed Chief of the NBS Statistical Engineering Laboratory in 1947 and in 1963 became a Senior Research Fellow. He retired from the NBS in 1983, but stayed on as a guest researcher. Churchill Eisenhart died from cancer at the age of 82 on 25 June 1994 [406, 1057, 1140].

In 1940 Abraham Wald (q.v. page 122) and Jacob Wolfowitz (q.v. page 122) published a new procedure to test whether two samples had been drawn from the same or identical populations [1405]. In this article they observed that in the problem treated by "Student," the distribution functions were assumed to be known, i.e., normal in form and completely specified by two parameters. They argued that such assumptions were open to very serious objections. For example, as they pointed out, the distributions may be radically different, yet have the same first moments. Wald and Wolfowitz proposed an alternative non-parametric procedure that was termed U and was based on the total number of runs and which, as Mood later stated, was very similar to his test published in the same year [999, p. 370].[15,16]

In 1943 Frieda Swed and Churchill Eisenhart, building on the runs test of Wald and Wolfowitz [1405], considered two different kinds of objects arranged along a

[14] The National Bureau of Standards was founded in 1901 as a non-regulatory agency of the United States Department of Commerce in Gaithersburg, Maryland. The NBS was renamed The National Institute of Standards and Technology (NIST) in 1988.

[15] In a 1943 article in *The Annals of Mathematical Statistics*, Wolfowitz commented that the choice of U as a test statistic was somewhat arbitrary and that other reasonable tests could certainly be devised [1465, p. 284], such as that proposed by Wilfrid Dixon in 1940, also in *The Annals of Mathematical Statistics* [353].

[16] A clear exposition of the Wald–Wolfowitz runs test was given in an article by Lincoln Moses on "Non-parametric statistics for psychological research" published in *Psychological Bulletin* in 1952 [1010].

line, e.g., $\{a, a, b, b, b, a, b\}$, where there are 3 as and 4 bs, forming four runs, two of as, i.e., $\{a, a\}$ and $\{a\}$ and two of bs, i.e., $\{b, b, b\}$ and $\{b\}$ [1337]. Thus, a run is a succession of similar events preceded and succeeded by different events; the number of elements in a run is referred to as its length [999, p. 367]. They showed that if there were m objects of one kind and n objects of the other kind, there were

$$\binom{m+n}{m}$$

equally-likely distinct arrangements of the objects under the null hypothesis, with $m \leq n$. If u is defined to be the number of distinct groupings of like objects, then the proportion of arrangements of the observed data yielding u' or fewer groupings is given by

$$P\{u \leq u'\} = \binom{m+n}{m}^{-1} \sum_{u=2}^{u'} f_u \ ,$$

where

$$f_u = 2\binom{m-1}{k-1}\binom{n-1}{k-1} \qquad \text{if } u = 2k$$

and

$$f_u = \left(\frac{m+n}{k} - 2\right)\binom{m-1}{k-1}\binom{n-1}{k-1} \qquad \text{if } u = 2k+1$$

for $1 \leq k \leq m \leq n$. Incidentally, the approximation to the normal distribution for u based on the large sample method of moments is given by

$$\frac{u - \mu_u}{\sigma_u} \ (\overset{d}{=})\ N(0, 1) \ ,$$

where the exact mean and variance of u are given by

$$\mu_u = \frac{2mn + m + n}{m+n}$$

and

$$\sigma_u^2 = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)} \ ,$$

respectively.

Swed and Eisenhart prepared three tables of probability values. The first table provided exact probability values for $P\{u \leq u'\}$ to seven places for $m \leq n \leq 20$ with a range of m from 2 to 20, inclusive. The second table provided exact probability values of u_ε for $\varepsilon = 0.005, 0.01, 0.025, 0.05, 0.95, 0.975, 0.99$, and 0.995, where u_ε is the largest integer, u', for which $P\{u \leq u'\} \leq \varepsilon$ when $\varepsilon < 0.50$ and is the smallest integer, u', for which $P\{u \leq u'\} \geq \varepsilon$ when $\varepsilon > 0.50$. The third table utilized the normal distribution provided by Wald and Wolfowitz, enhanced by a correction for continuity. In the third table, the values of u_ε from $m = n = 10$ through $m = n = 20$ were provided. There was a total of 736 values in the third table and all but five agreed with the exact values in the second table, leading Swed and Eisenhart to conclude "[i]t appears that the approximation will be adequate in general for $m = n \geq 20$" [1337, p. 67]. As Swed and Eisenhart noted, the merit of the test was that it employed a minimum of assumptions; only that the common population be continuous and the samples be independently drawn at random [1337]. The generalized runs test as presented by Mielke and Berry extended the Wald–Wolfowitz runs test from $t = 1$ tree, $g = 2$ groups, $N - k$ objects, and $L = N - 1$ links to $t \geq 1$ trees, $g \geq 1$ groups, and $N \geq k$ objects, where $N > k$ if $g = 1$ and $L \leq N - 1$ links [965, pp, 103–108].

3.6 Scheffé and Non-parametric Statistical Inference

In 1943 Henry Scheffé published an extensive 28 page introduction to statistical inference for non-parametric statistics in *The Annals of Mathematical Statistics*. At the time, this paper was considered a definitive work in the area of non-parametric tests and measures [1230].

H. Scheffé

Henry Scheffé received his A.B., A.M., and Ph.D. degrees in mathematics from the University of Wisconsin in 1931, 1933, and 1935, respectively. Scheffé held several academic positions at the University of Wisconsin, Oregon State University, Princeton University, Syracuse University, and the University of California at Los Angeles before accepting a position at Columbia University in 1948. In 1953 he moved to the University of California at Berkeley where he remained until his retirement in 1974. After his retirement from Berkeley, Scheffé accepted a 3-year appointment at the University of Indiana at Bloomington. Henry Scheffé returned to Berkeley in 1977 to complete work on a new edition of his magnum opus, *The Analysis of Variance*, but passed away 3 weeks later on 5 July at the age of 70, following a bicycling accident [814].

In 1943 Scheffé published what soon became a seminal article on non-parametric statistical inference. This article was an extensive introduction to what was then the relatively new field of non-parametric statistics [1230]. Scheffé prefaced this paper with an introduction in which he acknowledged that in most problems of statistical inference, where solutions do exist, the distribution function is assumed to depend on parameters, the values of which are unknown. Scheffé labeled this the "parametric case" under which, he said, falls all the theory based on normality assumptions [1230, p. 305]. He further observed that only a very small fraction of the extensive literature of mathematical statistics was devoted to the non-parametric case and that most of the non-parametric literature was quite recent. More formally, Scheffé defined a non-parametric test, noting that in any problem of statistical inference it is assumed that the cumulative distribution function F_n of the measurements is a member of a given class Ω of n-variate distribution functions. If Ω is a k-parameter family of functions the problem is called "parametric," otherwise it is called "non-parametric" [1230, p. 307].

In an extensive review and highly mathematical summary of the non-parametric literature, Scheffé provided an excellent description of permutation methods, which he termed "randomization methods" and attributed the origins of permutation methods to the work of R.A. Fisher in 1925 [448].[17] Scheffé noted that a special case of permutation methods was the "methods of ranks" to which he devoted considerable space and much detail. Near the end of a section on permutation methods, Scheffé mentioned a few difficulties with permutation methods when used in actual applications. The primary difficulty was, of course, that except for very small samples the calculation of exact permutation tests was "usually extremely tedious" [1230, p. 311]. He expressed dissatisfaction with those cases where the author of the test provided an approximation to the discrete permutation distribution by means of some familiar continuous distribution for which tables were readily available. He objected to "the laborious exact calculation by enumeration . . . being replaced by the computation of a few moments . . . and the use of existing tables of percentage points of the continuous distribution" [1230, p. 311]. Scheffé clearly took exception to the method of moments, emphasizing that with the exception of a few rank tests, the justification of these approximations was never satisfactory from a mathematical point of view, with the argument being based on two, three, or at most four moments.

In the following sections of this lengthy paper, Scheffé described in summary fashion the contributions to permutation tests by Hotelling and Pabst, who had previously investigated the Spearman coefficient of rank-order correlation [653]; Thompson and his rank test for two samples [1360], which was shown to be inconsistent with respect to certain alternatives by Wald and Wolfowitz [1405];

[17] Tests based on permutations of observations require that, under the null hypothesis, the probability distribution is symmetric under all permutations of the observations. This symmetry can be assured by randomly assigning treatments to the experimental units. As a result, these tests are often referred to as "randomization tests" in the literature [254, p. 729].

Swed and Eisenhart who provided tables for the 0.05 and 0.01 levels of significance for the runs test [1337]; Dixon who developed a two-sample test based on ranks for small sample sizes at the 0.01, 0.05, and 0.10 levels of significance [353]; Welch and his method of randomization for an analysis of variance ratio [1428]; Pitman's three randomization permutation tests for two independent samples, bivariate correlation, and randomized blocks analysis of variance [1129–1131]; Kendall's new τ measure of rank-order correlation with tables of exact probability values for small samples [728]; Kendall, Kendall, and Babington Smith and the permutation version of Spearman's rank-order correlation coefficient [746]; and Friedman and the analysis of variance for ranks [485], among others.

3.7 Wald–Wolfowitz and Serial Correlation

Early in the 1940s, Abraham Wald and Jacob Wolfowitz, both at Columbia University, published an exact permutation test of randomness based on serial correlation and designed for quality assurance analysis [1406].

A. Wald

Abraham Wald began his studies at the University of Cluj in Austria–Hungary (present-day Romania), then moved to the University of Vienna where he earned his Ph.D. in 1931. The pre-war environment provided few academic opportunities to Wald, who was Jewish, so he took a position as a tutor in mathematics. Wald immigrated to the United States in 1938, moving first to Colorado Springs, Colorado, to join the Cowles Commission for Research in Economics, but moving after only a few months to become a research associate at Columbia University at the invitation of Harold Hotelling. During his time at Columbia, Wald was a member of the Statistics Research Group (q.v. page 69) and it was while Wald was at Columbia that he met and began working with Jacob Wolfowitz, then a graduate student at Columbia. Abraham Wald remained at Columbia until his untimely death in a plane crash in southern India on 13 December 1950 at the age of 48 [814, 1426].

J. Wolfowitz

Jacob Wolfowitz earned his B.S. degree from City College of New York in 1931, his M.S. degree from Columbia University in 1933, and a Ph.D. from New York University in 1942. Between earning his M.S. and Ph.D. degrees, Wolfowitz taught high school mathematics to support his family, while continuing his studies at Columbia. Sometime in the 1930s, Wolfowitz

(continued)

began studying statistics at Columbia and it was at Columbia that Wolfowitz met Wald in 1938. In the spring of 1939, Wolfowitz and Wald began having long discussions about statistics, which resulted in Wald identifying a series of problems for the two of them to work on together. These discussions resulted in a series of collaborations in mathematical statistics and a lifelong friendship [814].

Wolfowitz also joined Wald as a member of the Statistics Research Group at Columbia in 1942 (q.v. page 69). Wolfowitz moved briefly to the University of North Carolina in 1945, but returned to Columbia University in 1946 where he remained until Wald's death in 1950.

In 1951 Wolfowitz took a position as professor of mathematics at Cornell University where he remained until 1970. In 1970 he joined the University of Illinois at Urbana, retiring in 1978, when he then went to the University of South Florida at Tampa as Shannon Lecturer in the Institute of Electrical and Electronic Engineers. Jacob Wolfowitz suffered a heart attack and passed away on 16 July 1981 at the age of 71 in Tampa, Florida [9].

As an interesting aside, Wolfowitz has been credited [323] with coining the term "non-parametric" in his 1942 paper on "Additive partition functions and a class of statistical hypotheses" in *The Annals of Mathematical Statistics* [1464, p. 264].

In 1943 Wald and Wolfowitz devised an exact non-parametric test for randomness based on serial correlation [1406]. Noting that the problem of testing randomness frequently arose in quality control of manufactured products and in the analysis of time series data, Wald and Wolfowitz constructed an exact permutation test based on serial correlation with a defined lag [1406]. Following the notation of Wald and Wolfowitz, suppose that x denotes some quality character of a product and that $x_1, \ldots, x_N$ are the values of x for N consecutive units of the product arranged in the order they were produced. The production process is considered to be in a state of statistical control if the sequence $\{x_1, \ldots, x_N\}$ satisfies the condition of randomness. The serial correlation coefficient with lag h is defined as

$$R_h = \frac{\displaystyle\sum_{i=1}^{N} x_i\, x_{h+i} - \left(\sum_{i=1}^{N} x_i\right)^2 \Bigg/ N}{\displaystyle\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 \Bigg/ N}\,, \qquad (3.2)$$

where x_{h+i} is to be replaced by x_{h+i-N} for all values of i for which $h + i > N$.

Denote by a_i the observed value of x_i, $i = 1, \ldots, N$, and consider the subpopulation where the set $\{x_1, \ldots, x_N\}$ is restricted to permutations of $a_1, \ldots, a_N$. In the subpopulation, the probability that $\{x_1, \ldots, x_N\}$ is any particular permutation

$\{a'_1, \ldots, a'_N\}$ of $\{a_1, \ldots, a_N\}$ is equal to $1/N!$ if the hypothesis of randomness is true. Then the probability distribution of R_h in the subpopulation can be determined. Consider the set of $N!$ values of R_h obtained by substituting for $\{x_1, \ldots, x_N\}$ all possible permutations of $\{a_1, \ldots, a_N\}$. Each value of R_h has the probability $1/N!$. Let α denote the level of significance and choose as a critical region a subset of M values of the set of $N!$ values of R_h where $M/N! = \alpha$.

Next, Wald and Wolfowitz considered the statistic

$$\bar{R}_h = \sum_{i=1}^{N} x_i \, x_{h+i} \ ,$$

where x_{h+i} is to be replaced by x_{h+i-N} for all values of i for which $h + i > N$. They pointed out that since $\sum_{i=1}^{N} x_i$ and $\sum_{i=1}^{N} x_i^2$ in Eq. (3.2) are constants and therefore invariant under permutation, the statistic $\bar{R}_h$ is a linear function of R_h in the subpopulation and could be substituted for R_h when constructing the exact permutation distribution [1406]. This dramatically simplified calculations. However, as Wald and Wolfowitz noted, difficulties in carrying out the test arose if N was neither sufficiently small to make the computations of the $N!$ values of $\bar{R}$ practically possible, nor sufficiently large to permit the use of a limiting distribution. They concluded that "it may be helpful to determine the third, fourth, and perhaps higher, moments of $\bar{R}$, on the basis of which upper and lower limits for the cumulative distribution of $\bar{R}$ can be derived" [1406, p. 381]. The remainder of the paper is devoted to deriving the mean and variance of $\bar{R}$. Finally, Wald and Wolfowitz observed that they could replace the observed values $\{a_1, \ldots, a_N\}$ by their ranks, but questioned the wisdom in making the test on the rank-transformed values instead of the original observations [1406, p. 387].

The following year Wald and Wolfowitz published a general paper on a variety of statistical tests based on permutations of observations [1407]. They observed that one of the problems of statistical inference was to devise exact tests of significance when the form of the underlying probability distribution was unknown, such as Fisher had discussed in 1925 and 1935 [448, 451]. They explained that an exact test on the level of significance α could be constructed by choosing a proportion α of the permutations of the observations as a critical region. Wald and Wolfowitz noted that Scheffé had previously shown that for a general class of problems this was the only possible method of constructing exact tests of significance [1230].

In this 1944 paper, a general theorem on the limiting distribution of linear forms in the universe of permutations of the observations was derived by Wald and Wolfowitz. Applications of this general theorem were made by Wald and Wolfowitz for a number of existing permutation tests, including Spearman's rank-order correlation coefficient [1300, 1301], the limiting distribution which had previously been proved by Hotelling and Pabst [653]; Pitman's test of dependence between two variates [1130]; Pitman's test of the hypothesis that two samples came from the same population [1129]; the analysis of variance for randomized blocks as developed by both Pitman and Welch [1131, 1428]; and Hotelling's generalized T^2

for permutations of the observations [652]. In this last case, Wald and Wolfowitz derived the limiting distribution of T^2 in the universe of permutations of the observations, an original contribution by Wald and Wolfowitz to the Hotelling paper [652].

3.8 Mann and a Test of Randomness Against Trend

In 1945 Henry Mann introduced a two-sample rank test of randomness based on the number of reversal arrangements necessary to convert one set of ranks into a second set of ranks [879].

H.B. Mann

Henry Berthold Mann received his Ph.D. in mathematics from the University of Vienna in 1935, then emigrated from Austria to the United States in 1938. In 1942 he was the recipient of a Carnegie Fellowship for the study of statistics at Columbia University where he had the opportunity to work with Abraham Wald in the Department of Economics, which at the time was headed by Harold Hotelling. This likely would have put him in contact with other members of the Statistical Research Group at Columbia University such as W. Allen Wallis, Jacob Wolfowitz, Milton Friedman, Jimmie Savage, Frederick Mosteller, and Churchill Eisenhart (q.v. page 69).

In 1946 Mann accepted a position at The Ohio State University, remaining there until his retirement in 1964, at which point he moved to the U.S. Army's Mathematics Research Center at the University of Wisconsin. In 1971, Mann moved again to the University of Arizona, retiring a second time in 1975. Henry Mann remained in Arizona until his death on 1 February 2000 at the age of 94 [1060].

In 1945 Mann introduced two non-parametric tests of randomness against trend [879]. Both tests were based on ranks, but it is the first of his two tests that pertains to permutation statistical methods. Mann noted, as had others, that the advantages of such rank tests are that they may be used if the quantities considered cannot be measured, as long as it is possible to rank the observations [879, p. 247]. By way of examples of such quantities, Mann specifically mentioned ranking the intensity of sensory impressions, such as pleasure and pain.

As Mann explained, let $X_{i_1}, \ldots, X_{i_n}$ be a permutation of the n distinct numbers $X_1, \ldots, X_n$, let T count the number of inequalities $X_{i_k} < X_{i_l}$ where $k < l$, and label one such inequality a "reverse arrangement." If $X_1, \ldots, X_n$ all have the same continuous distribution, then the probability of obtaining a sample of size n with T reversal arrangements is proportional to the number of permutations of the variables $1, 2, \ldots, n$ with T reversal arrangements. Mann stated that the statistic T was first

Table 3.1 Reversal sequences for $n = 5$ ranks to obtain no reversals from an observed data set

		Reversal sequence													
Observed		1		2		3		4		5		6		7	
1	3	1	3	1	3	1	3	1	1	1	1	1	1	1	1
2	4	2	4	2	4	2	1	2	3	2	3	2	3	2	2
3	5	3	5	3	1	3	4	3	4	3	4	3	2	3	3
4	2	4	1	4	5	4	5	4	5	4	2	4	4	4	4
5	1	5	2	5	2	5	2	5	2	5	5	5	5	5	5

proposed by M.G. Kendall in 1938 and acknowledged that Kendall had also derived a recursion formula, tabulated the distribution of T for $T \leq 10$, and proved that the asymptotic distribution of T is normal; see Sect. 2.18 in Chap. 2. What Mann contributed in his paper was a table of probability values that was easier to use and a simpler proof of the normality of the asymptotic distribution of T. The table produced by Mann provided cumulative probability values for $3 \leq n \leq 10$ with T or fewer reversal arrangements, where $0 \leq T \leq 21$ and every permutation occured with probability $1/n!$. The counting of the reversal arrangements followed the technique described by M.G. Kendall in 1938 [728].

Table 3.1 illustrates the counting of reversal arrangements in a sequence of ranks from 1 to 5. The first set of two columns in Table 3.1 lists the observed ranks for two groups, and subsequent sets of columns illustrate the number of reversals necessary to produce the first column from the second. In this case, seven reversal sequences are required with one reversal arrangement per sequence. For example, reversal sequence 1 in Table 3.1 exchanges ranks 2 and 1 in the observed column, reversal sequence 2 exchanges ranks 5 and 1 in reversal sequence 1, reversal sequence 3 exchanges ranks 4 and 1 in reversal sequence 2, and so on until reversal sequence 7 exchanges ranks 3 and 2 in reversal sequence 6 to achieve the ordered sequence in reversal sequence 7.

The technique that Mann described is similar to a graphic computation of disarray first constructed by S.D. Holmes and published in an appendix to a book on *Educational Psychology* by P. Sandiford in 1928 with application to the Pearson product-moment correlation coefficient [1221, pp. 391–394], and in a later publication by H.D. Griffin in 1958 with reference to the Kendall rank-order correlation coefficient, τ_a [558].

A proof that the number of interchanges of nearest neighbors required to reduce one ranking to the other is related to T was provided by P.A.P. Moran in 1947 [1003] and was, according to Moran, first proved by Olinde Rodrigues in 1839 [1182].[18] On this note, in 1948 Moran mathematically established the relationship between

[18] A summary in English of the Rodrigues 1839 article is available in *Mathematics and Social Utopias in France: Olinde Rodrigues and His Times* [39, pp. 110–112].

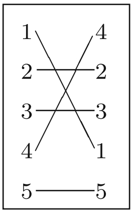


Fig. 3.1 Graphic depiction of the number of reversals for two sets of ranks, from 1 to 5

rank-order correlation and permutation distributions [1004].[19] Consider n objects denoted by $1, \ldots, n$ and let s be the least number of interchanges of adjacent objects required to restore the permutations to the normal order. In his 1938 article that introduced a new coefficient of rank-order correlation, τ, Kendall (q.v. page 84) showed that $S = \tau n(n-1)/2$ is distributed about a mean of zero with variance given by $n(n-1)(2n+5)/18$ in a distribution that tended to normality as n increased [728]. Utilizing a theorem of Haden [569], Moran proved that $s = n(n-1)/4 - S/2$ so that

$$\tau = 1 - \frac{4s}{n(n-1)} = -\frac{4t}{n(n-1)} \ ,$$

where $t = s - n(n-1)/4$. This showed that Kendall's τ rank-order correlation coefficient could be defined in terms of s and, therefore, the theory of rank-order correlation could be mathematically linked with the theory of permutations. This ultimately became an observation of considerable importance.

A graphic that depicts the number of reversals consists of lines that are drawn between like values in the two columns and the number of reversals is represented by the number of times the lines cross [558]. For example, consider the two sets of ranks given in Fig. 3.1.[20]

There are five crosses (×) among the $n = 5$ lines, i.e., both diagonal lines cross two horizontal lines and each other, indicating the five reversals required to produce the distribution of ranks on the left from the distribution of ranks on the right. Thus, beginning with the right column of {4, 2, 3, 1, 5} and for the first reversal, exchange ranks 3 and 1, yielding {4, 2, 1, 3, 5}; for the second reversal, exchange ranks 2 and 1, yielding {4, 1, 2, 3, 5}; for the third reversal, exchange ranks 4 and 1, yielding {1, 4, 2, 3, 5}; for the fourth reversal, exchange ranks 4

[19] This paper was cited by Moran in [1005, p. 162] as "Rank correlation and a paper by H.G. Haden," but apparently the title was changed at some point to "Rank correlation and permutation distributions" when it was published in *Proceedings of the Cambridge Philosophical Society* in 1948.

[20] Technically, Fig. 3.1 is a permutation graph of a family of line segments that connect two parallel lines in the Euclidean plane. Given a permutation $\{4, 2, 3, 1, 5\}$ of the positive integers $\{1, 2, 3, 4, 5\}$, there exists a vertex for each number $\{1, 2, 3, 4, 5\}$ and an edge between two numbers where the segments cross in the permutation diagram.

Table 3.2 Permutations and number of reversals for $n = 4$ ranks: $\{1, 2, 3, 4\}$

Number	Permutation	Reversals	Number	Permutation	Reversals
1	1 2 3 4	0	13	3 1 2 4	2
2	1 2 4 3	1	14	3 1 4 2	3
3	1 3 2 4	1	15	3 2 1 4	3
4	1 3 4 2	2	16	3 2 4 1	4
5	1 4 2 3	2	17	3 4 1 2	5
6	1 4 3 2	3	18	3 4 2 1	3
7	2 1 3 4	1	19	4 1 2 3	4
8	2 1 4 3	2	20	4 1 3 2	4
9	2 3 1 4	2	21	4 2 1 3	4
10	2 3 4 1	3	22	4 2 3 1	5
11	2 4 1 3	3	23	4 3 1 2	5
12	2 4 3 1	4	24	4 3 2 1	6

and 2, yielding {1, 2, 4, 3, 5}; and for the fifth reversal, exchange ranks 4 and 3, yielding {1, 2, 3, 4, 5}.

To illustrate the Mann procedure to obtain exact probability values under the null hypothesis of randomness, $P(X_i > X_k) = 1/2$, consider an example with $n = 4$ ranks, where there are $n! = 4! = 24$ possible permutations of the ranks. Table 3.2 lists the 24 permutations of the four ranks, along with the number of reversal arrangements required to achieve a sequence of {1, 2, 3, 4}. As can be seen in Table 3.2, there is only one permutation with no (zero) reversal arrangements, i.e., permutation number 1 with {1, 2, 3, 4}. Thus, the probability of zero reversal arrangements is

$$P(T = 0) = \frac{1}{24} = 0.0417\,.$$

There are three permutations with one reversal arrangement, i.e., permutations 2, 3, and 7; thus, the cumulative probability of one or fewer reversal arrangements is

$$P(T \leq 1) = \frac{3}{24} + \frac{1}{24} = \frac{4}{24} = 0.1667\,.$$

There are five permutations with two reversal arrangements, i.e., permutations 4, 5, 8, 9, and 13; thus, the cumulative probability of two or fewer reversal arrangements is

$$P(T \leq 2) = \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{9}{24} = 0.3750\,.$$

There are six permutations with three reversal arrangements, i.e., permutations 6, 10, 11, 14, 15, and 18; thus, the cumulative probability of three or fewer reversal arrangements is

Fig. 3.2 Portion of a figurate triangle with $n = 4$ and 5

n	Partial figurate triangle										
4	1	3	5	6	5	3	1				
5	1	4	9	15	20	22	20	15	9	4	1

$$P(T \leq 3) = \frac{6}{24} + \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{15}{24} = 0.6250 \,.$$

There are five permutations with four reversal arrangements, i.e., permutations 12, 16, 19, 20, and 21; thus, the cumulative probability of four or fewer reversal arrangements is

$$P(T \leq 4) = \frac{5}{24} + \frac{6}{24} + \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{20}{24} = 0.8333 \,.$$

There are three permutations with five reversal arrangements, i.e., permutations 17, 22, and 23; thus, the cumulative probability of five or fewer reversal arrangements is

$$P(T \leq 5) = \frac{3}{24} + \frac{5}{24} + \frac{6}{24} + \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{23}{24} = 0.9583 \,.$$

Finally, there is only one permutation with six reversal arrangements, i.e., permutation 24 with $\{4, 3, 2, 1\}$; thus, the cumulative probability of six or fewer reversal arrangements is

$$P(T \leq 6) = \frac{1}{24} + \frac{3}{24} + \frac{5}{24} + \frac{6}{24} + \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{24}{24} = 1.0000 \,.$$

Note that in this example with $n = 4$ ranks, the numerators of the final fractions are 1, 3, 5, 6, 5, 3, 1 and, using the Kendall recursion procedure (q.v. page 86), Mann was able to generate numerator values for successive values of n up to 10. For example, consider $n = 5$ where the denominator is $n! = 5! = 120$ and there are 11 numerator values instead of 7. The process is as follows. First for $n = 2$ there are two numerator values, for $n = 3$ there are four values, for $n = 4$ there are seven values, and for $n = 5$ there are 11 numerator values. Thus, add $n - 1$ to the previous number of values, e.g., for $n = 2$ with two values, $n = 3$ will have $2 + n - 1 = 2 + 3 - 1 = 4$ values, $n = 4$ will have $7 + n - 1 = 7 + 5 - 1 = 11$ values and so on. The numerator values for $n = 5$ can be obtained from the numerator values for $n = 4$ by a figurate triangle, a portion of which is listed in Fig. 3.2. For the full figurate triangle, see page 86 in Chap. 2.

Here, as in Kendall's 1938 article [728], a number in the nth row is the sum of the number immediately above it and the $n - 1$ or fewer numbers to the left of that number, e.g., in row $n = 5$ the number 9 in the third position from the left is the sum of the number above it (5) in row 4 and all the numbers to the left of 5 in row 4 (3 and 1), since there are fewer than $n - 1 = 5 - 1 = 4$ numbers to the left of 3; and in row $n = 5$, 22 in the sixth position from the left is the sum of the number above it (3) and the $n - 1 = 5 - 1 = 4$ numbers to the left of 3: $22 = 3 + 5 + 6 + 5 + 3$.

In this manner, Mann constructed a table of exact probability values for a test of trend with $3 \leq n \leq 10$, under the null hypothesis of randomness. The remainder of the article was concerned with determining approximate probability values for T. Under the null hypothesis of randomness Mann obtained the mean of T, i.e., $E(T) = n(n-1)/4$, and continued to find the higher moments beyond the mean. He then proved that the limiting distribution of T was normal.

Gottfried Emanuel Noether further investigated certain asymptotic properties of the test of randomness based on the statistic R_h proposed by Wald and Wolfowitz [1038]. He was able to show that the conditions given in the original paper by Wald and Wolfowitz [1406] for the asymptotic normality of R_h when the null hypothesis of randomness was true could be weakened considerably. Further, Noether described a simple condition for the asymptotic normality of R_h for ranks under the alternative hypothesis. He then utilized this asymptotic normality to compare the asymptotic power of R_h with the T statistic proposed by Mann [879] in the case of downward trend [1038].

3.9 Barnard and 2 × 2 Contingency Tables

In 1945 George Barnard introduced the CSM test for 2 × 2 contingency tables that was based on two binomial distributions representing the two rows of the observed contingency table [63]. In this article, Barnard claimed that the proposed CSM test was more powerful than the Fisher–Yates exact probability test.

G.A. Barnard

George Alfred Barnard attended St. John's College, University of Cambridge, on a scholarship, earning a degree in mathematics in 1937. From 1937 to 1939 he did graduate work on mathematical logic under Alonzo Church at Princeton University. Another prominent Englishman who was at Princeton at the same time as Barnard and also studying under Church was Alan Mathison Turing, British logician, cryptologist, and the "father of computer science and artificial intelligence" [188, p. 272].[21]

Barnard was on holiday in Great Britain when World War II began and he never returned to Princeton to finish his Ph.D. Complicating the matter was Barnard's radical left-wing views, a consequence of which was that he was denied a visa for the United States for many years after the war; for Barnard's views on this, see "A conversation with George A. Barnard" by Morris DeGroot published in *Statistical Science* in 1988 [339, p. 206].

(continued)

[21] See also a discussion about the relationship between Turing and Church by George Dyson in a 2012 book titled *Turing's Cathedral* [370, pp. 249–250].

As Dennis Lindley noted, Barnard's expressions of his anti-establishment views most likely accounts for Barnard never being elected to a Fellowship in the Royal Society [831].

In 1940 Barnard accepted a position at the Plessey Company, an engineering firm, as a mathematical consultant, and in 1942 he joined the Ministry of Supply as head of a research group that applied quality control to the products for which they were responsible. It was at that time that he developed an interest in statistics. The research group that he supervised included Dennis Lindley; Peter Armitage; Robin Plackett; Peter Burman; Patrick Rivett, who subsequently went into operational research as the first professor of operational research in the United Kingdom; Dennis Newman, of the Newman–Keuls test; and Frank Anscombe [339].

At the conclusion of the war, Barnard accepted an appointment at Imperial College, London, where he was named professor of mathematics in 1954, but he left in 1966 for the newly created University of Essex, from which he retired in 1975 (the University of Essex in Colchester was established in 1963 and received its Royal Charter in 1965). After retirement, Barnard spent much of each year, until 1981, at the University of Waterloo in Ontario, Canada. George Alfred Barnard died peacefully in Brightlingsea, Essex, on 30 July 2002 at the age of 86 [830, 831]. For some personal insights on the life of George Barnard, see the fourth chapter in G.E.P. Box's autobiography *An Accidental Statistician* published in 2013 [192, Chap. 4].

In 1945 George Barnard introduced a new test for 2 × 2 contingency tables that he claimed was more powerful than the Fisher–Yates exact probability test [63].[22] Taking the table to be generated by samples of n_1 and n_2 from two binomial distributions with probabilities p_1 and p_2, respectively, Barnard argued that if $p_1 = p_2 = p$ and $n_1 = n_2 = 3$, for example, the probability of observing a 2×2 contingency table with rows $\{3, 0\}$ and $\{0, 3\}$ was $p^3(1 - p)^3$, which gave the probability value 1/64 when $p = 0.5$ and was less than this for all other values of p, as opposed to a probability value of 1/20 if all marginal frequency totals were regarded as fixed, as Fisher had recommended.

The new test prompted an exchange between Fisher and Barnard [64, 457], debating the merits of both methods; see also articles by Barnard in 1947 and 1949 [67, 68] and by E.S. Pearson in 1947 [1095]. In 1947 Barnard named the test the CSM test [67, p. 124], but in 1949, in a paper read before the Research Section of the Royal Statistical Society, Barnard withdrew the test from further consideration. He allowed as he had never been satisfied with the position he had taken in 1945 and said that "further meditation has led me to think that Professor Fisher was right

[22] This was actually Barnard's first, of many, published papers. It was published in *Nature* while Barnard was employed at the Ministry of Supply and is only one-half page in length.

after all" [68, p. 115]. He credited Egon Pearson for strengthening this conclusion by his remarks in his article on choosing statistical tests [1095].[23,24]

3.10 Wilcoxon and the Two-Sample Rank-Sum Test

Frank Wilcoxon, trained as a chemist, was also an accomplished statistician. In 1945 Wilcoxon, in a concise article in the first volume of *Biometrics Bulletin*, introduced two new rank tests: the two-sample rank-sum test for two independent (unpaired) samples, and the matched-pairs (signed-ranks) rank-sum test for two dependent (paired) samples [1453].

F. Wilcoxon

Frank Wilcoxon had an interesting early life. Wilcoxon's parents were wealthy Americans and were honeymooning in Europe. They rented the Glengarriff Castle near Cork, Ireland, where Wilcoxon and his twin sister were born on 2 September 1892. In 1908, at the age of 16, Wilcoxon ran away to sea. At some point he jumped ship and hid for years in the back country of West Virginia working as an oil-well worker and a tree surgeon [221]. Returning home to Catskill, New York, he enrolled at the Pennsylvania Military College. Wilcoxon earned his B.Sc. degree from Pennsylvania Military College in 1917, an M.S. degree in chemistry from Rutgers University in 1921, and a Ph.D. in chemistry from Cornell University in 1924.

Wilcoxon spent much of his adult life as a chemist working for the Boyce Thompson Institute for Plant Research in Yonkers, New York, the Atlas Powder Company in Wilmington, Delaware, and, finally, the Lederle Laboratories Division at the American Cyanamid Company in Norwalk, Connecticut. It was while at the Boyce Thompson Institute that Wilcoxon's interest in statistics was spurred through his work with a small reading group that met to study R.A. Fisher's *Statistical Methods for Research Workers*. Organizers of the group were Wilcoxon, fellow chemist William John (Jack) Youden, and biologist Frank E. Denny. This introduction to statistics had a

(continued)

[23] In 1984 Barnard revealed the meaning behind labeling the statistic CSM, recalling "there was a private pun in my labelling the suggested procedure CSM—it referred . . . to the Company Sergeant Major in my Home Guard unit at the time, my relations with whom were not altogether cordial. I still feel that the test, like the man, is best forgotten" [70, p. 450].

[24] The Barnard test will not die and from time to time the test is resurrected and advocated; see for example, articles by McDonald, Davis, and Milliken [913] in 1977; Barnard [72], Hill [621], and Rice [1167, 1168] in 1988; Dupont [365] and Martín Andrés and Luna del Castillo [900] in 1989; and Campbell [239] in 2007.

profound effect on the subsequent careers of Wilcoxon and Youden as both became leading statisticians of the time. Wilcoxon retired from the American Cyanamid Company in 1957 and 3 years later, at the behest of Ralph Bradley, joined the faculty at Florida State University in Tallahassee, Florida, where he helped develop its Department of Statistics.

As Bradley related, he and Wilcoxon had met several times at Gordon Research Conferences,[25] and in 1959 Bradley was recruited from Virginia Polytechnic Institute to initiate a department of statistics at Florida State University (formerly, the Florida State College for Women). Bradley persuaded Wilcoxon, who had retired in Florida, to come out of retirement and join the newly-formed department. Wilcoxon agreed to a half-time position teaching applied statistics as he wanted time off to kayak and ride his motorcycle [639]. Frank Wilcoxon died on 18 November 1965 after a brief illness at the age of 73 [203–205]. At the time of his death, Wilcoxon was Distinguished Lecturer in the Department of Statistics at Florida State University [363].

In 1945 Wilcoxon introduced a two-sample test statistic, W, for rank-order statistics [1453].[26] In this very brief paper of only three pages Wilcoxon considered the case of two samples of equal sizes and provided a table of exact probability values for the lesser of the two sums of ranks for both paired and unpaired experiments [1453]. In the case of unpaired samples, a table provided exact probability values for 5–10 replicates in each sample; and for paired samples, a table provided exact probability values for 7–16 paired comparisons.[27] Bradley has referred to the unpaired and paired rank tests as the catalysts for the flourishing of non-parametric statistics [639] and Brooks described the Wilcoxon 1945 article as “a bombshell which broke new and permanent ground” and the unpaired and paired rank tests as “cornerstones in the edifice of nonparametric statistics” [221].

[25] The Gordon Research Conferences on Statistics in Chemistry and Chemical Engineering began in 1951 and continued through the summer of 2005.

[26] The Wilcoxon two-sample rank-sum test statistic is conventionally expressed as W in textbooks, but Wilcoxon actually designated his test statistic as T. Also, many textbooks describe the Wilcoxon test as a “difference between group medians” test, when it is clearly a test for the difference between mean ranks; see for example, an article by Bergmann, Ludbrook, and Spooren in 2000 [100] and an article by Conroy in 2012 [274].

[27] A clear and concise exposition of the Wilcoxon unpaired and paired sample rank tests is given in an article by Lincoln Moses on “Non-parametric statistics for psychological research” published in *Psychological Bulletin* in 1952 [1010].

Wilcoxon's 1945 Article

While Wilcoxon's 1945 article may have been a "bombshell," it should be emphasized that the original Wilcoxon article can be difficult to read. It is cryptic, incomplete, and contains notational errors. Moreover, Wilcoxon was trained as a chemist and not as a mathematician; consequently, his notation was somewhat unconventional. For example, as late as 1945 Wilcoxon was still using the old representation for n factorial of $\lfloor\underline{n}$ instead of the customary $n!$ expression. The expression $\lfloor\underline{n}$ was developed by Thomas Jarrett, an English churchman and orientalist, and first published in 1830, appearing in a paper "On algebraic notation" printed in *Transactions of the Cambridge Philosophical Society* [681, p. 67]. The familiar $n!$ expression was introduced by the French mathematician Chrétien (Christian) Kramp as a convenience to his printer who was unable to typeset Jarrett's $\lfloor\underline{n}$. The factorial symbol $n!$ first appeared in Kramp's book on *Éléments d'arithmétique universelle* in 1808 [770].[28]

3.10.1 Unpaired Samples

Wilcoxon showed that in the case of two unpaired samples with rank numbers from 1 to $2q$, where q denotes the number of ranks (replicates) in each sample, the minimum sum of ranks possible is given by $q(q + 1)/2$, where W is the sum of ranks in one sample, continuing by steps up to the maximum sum of ranks given by $q(3q + 1)/2$. For example, consider two samples of $q = 5$ measurements converted to ranks from 1 to $2q = 10$. The minimum sum of ranks for either group is $\{1 + 2 + 3 + 4 + 5\} = 5(5 + 1)/2 = 15$ and the maximum sum of ranks is $\{6 + 7 + 8 + 9 + 10\} = 5[(3)(5) + 1]/2 = 40$. Wilcoxon explained that these two values could be obtained in only one way, but intermediate sums could be obtained in more than one way. For example, the sum of $T = 20$ could be obtained in seven ways, with no part greater than $2q = 10$: $\{1, 2, 3, 4, 10\}$, $\{1, 2, 3, 5, 9\}$, $\{1, 2, 3, 6, 8\}$, $\{1, 2, 4, 5, 8\}$, $\{1, 2, 4, 6, 7\}$, $\{1, 3, 4, 5, 7\}$, and $\{2, 3, 4, 5, 6\}$. The number of ways each sum could arise is given by the number of q-part, here 5-part partitions of $T = 20$, the sum in question.[29]

[28] For a brief history of the factorial symbol, see a 1921 article in *Isis* by Florian Cajori on the "History of symbols for $\underline{\text{n}}$ = factorial" [237].

[29] Wilcoxon's use of the term "partitions" here is a little misleading. These are actually sums of $T = 20$, each sum consisting of five integer values between 1 and $2q = 10$ with no integer value repeated e.g., $\{1, 2, 3, 4, 10\} = 20$ which consists of five non-repeating integer values, but not $\{5, 7, 8\} = 20$ which consists of only three integer values, nor $\{1, 3, 3, 5, 8\} = 20$ which contains multiple values of 3.

This was not a trivial problem to solve, as calculating the number of partitions is quite difficult, even today with the availability of high-speed computers. In general, the problem is known as the "subset-sum problem" and requires a generating function to solve. The difficulty is in finding all subsets of a set of numbers that sum to a specified total, without repetitions. The approach that Wilcoxon took was ingenious and is worth examining, as the technique became the basic method for other researchers as well as the basis for several computer algorithms in later years. Wilcoxon showed that the required partitions were "equinumerous" with another set of partitions, r, that were much easier to enumerate, a technique he apparently learned from a volume by Percy Alexander MacMahon on *Combinatory Analysis* [865].[30] He defined r as the serial number of T in the possible series of sums, beginning with 0, i.e., $0, 1, 2, \ldots, r$.

For an illustrative example, consider *vide supra* $q = 5$ replications of measurements on two samples and assign ranks 1 through $2q = 10$ to the data: $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. The lowest possible sum is $1+2+3+4+5 = 15$ and the highest possible sum is $6+7+8+9+10 = 40$. Then the question is: in how many ways can a total of $T = 20$ be obtained, i.e., how many unequal five-part partitions of $T = 20$ exist, having no part greater than $2q = 10$ and no repetition of values? As shown above, there are seven such partitions. Now, $T = 20$ is sixth in the possible series of totals, i.e., $T = 15, 16, 17, 18, 19, 20, \ldots, 40$; therefore, $r = 5$ and the total number of partitions that sum to $T = 20$ is equivalent to the total number of partitions that sum to $r = 5$ with no part greater than $q = 5$; specifically, $\{5\}$, $\{1, 4\}$, $\{2, 3\}$, $\{1, 1, 3\}$, $\{1, 2, 2\}$, $\{1, 1, 1, 2\}$, and $\{1, 1, 1, 1, 1\}$. These are, of course, true partitions consisting of one to five integer values between 1 and 5, summing to 5 with repetitions allowed. Wilcoxon capitalized on the relationship between the two subset-sum problems, $T = 20$ and $r = 5$, to enumerate the partitions of $r = 5$ from an available table of partitions included in a 1942 book by William Allen Whitworth titled *Choice and Chance* [1446], which then corresponded to the more difficult enumeration of the five-part partitions of $T = 20$.

Partitions

The number theory function known as the partition function gives the number of ways of writing any integer as a sum of smaller positive integers [505]. As above, the integer 5 can be written in seven different ways: $\{5\}$, $\{1, 4\}$, $\{2, 3\}$, $\{1, 1, 3\}$, $\{1, 2, 2\}$, $\{1, 1, 1, 2\}$, and $\{1, 1, 1, 1, 1\}$. The partition number of

(continued)

[30] MacMahon's monumental two-volume work on *Combinatory Analysis*, published in 1916, contained a section in Volume II, Chap. III, on "Ramanujan's Identities" in which MacMahon demonstrated the relationship between the number of q-part unequal partitions without repetitions with no part greater than $2q$ and the number of partitions with repetitions with no part greater than q [865, pp. 33–48].

Fig. 3.3 Values of T and corresponding values of r

T:	28,	29,	30,	31,	32,	33,	34,	35,	36,	...,	77
r:	0,	1,	2,	3,	4,	5,	6,	7,	8,	...,	49

5 is, therefore, 7. As given by Gelman [505, p. 183], the partition numbers of the integers from 1 to 10 are:

Integer:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Partition number:	1, 2, 3, 5, 7, 11, 15, 22, 30, 42

Thus, there are 30 ways to sum smaller integers to make a sum of 9 and 42 ways to sum smaller integers to make a sum of 10, with repetitions. As Gelman observed, while the partition number of 100 is only 190,569,292, the partition number of 1,000 is an astounding 24,061,467,864,032,622,473,692,149,727,991 [505, p. 183]. In 1918 Godfrey Harold (G.H.) Hardy and Srinivasa Ramanujan, in a remarkable article in *Proceedings of the London Mathematical Society*, provided the asymptotic formula for partition numbers, $p(n)$, showing that as $n \to \infty$,

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right)$$

[585, p. 79]. See also a 2004 article on this topic by Berry, Johnston, and Mielke in *Psychological Reports* [111].

To illustrate the Wilcoxon procedure, consider an example two-sample rank-sum test analysis with $q = 7$ replicates in each treatment and the lesser of the two sums of ranks $T = 35$. The minimum value of T with $q = 7$ replicates is $T = 1 + 2 + 3 + 4 + 5 + 6 + 7 = 7(7+1)/2 = 28$. The values of T with the corresponding values of r are given in Fig. 3.3.

The exact lower one-sided probability (P) value of $T = 35$ is given by

$$P = \left\{ 1 + \sum_{i=1}^{r} \sum_{j=1}^{q} \mathbb{P}_j^i - \sum_{k=1}^{r-q} \left[(r - q - k + 1)\mathbb{P}_{q-1}^{q-2+k} \right] \right\} \bigg/ \frac{(2q)!}{(q!)^2} \ ,$$

where $\mathbb{P}_j^i$ represents the number of j-part partitions of i; r is the serial number of possible rank totals, 0, 1, 2, ..., r; and q is the number of replicates [1453, p. 82]. If $q \geq r$, the summation $\sum_{k=1}^{r-q}$ is assumed to be zero. For the example data, the equation is

$$P = \left\{ 1 + \sum_{i=1}^{7} \sum_{j=1}^{7} \mathbb{P}_j^i - \sum_{k=1}^{7-7} \left[(7 - 7 - k + 1)\mathbb{P}_{7-1}^{7-2+k} \right] \right\} \bigg/ \frac{14!}{(7!)^2} \ ,$$

Table 3.3 Illustrative table comparing the $q = 7$-part partitions of $T = 35$ with the corresponding partitions of $r = 7$

	Partition													
Number	$q = 7,\ T = 35$							$r = 7$						
1	1	2	3	4	5	6	14	1	1	1	1	1	1	1
2	1	2	3	4	5	7	13		1	1	1	1	1	2
3	1	2	3	4	5	8	12			1	1	1	2	2
4	1	2	3	4	6	7	12				1	2	2	2
5	1	2	3	4	5	9	11			1	1	1	1	3
6	1	2	3	4	6	8	11				1	1	2	3
7	1	2	3	5	6	7	11					2	2	3
8	1	2	3	4	6	9	10					1	3	3
9	1	2	3	4	7	8	10				1	1	1	4
10	1	2	3	5	6	8	10					1	2	4
11	1	2	4	5	6	7	10						3	4
12	1	2	3	5	7	8	9					1	1	5
13	1	2	4	5	6	8	9						2	5
14	1	3	4	5	6	7	9						1	6
15	2	3	4	5	6	7	8							7

and the exact lower one-sided probability value is

$$P = \{1 + 1 + 2 + 3 + 5 + 7 + 11 + 15 - 0\} / \left[87{,}178{,}291{,}200/(5{,}040^2\right]$$
$$= 45/3{,}432 = 0.0131 \ .$$

The correspondence between the number of unequal q-part partitions of T with no part greater than $2q$ and the number of partitions of r with no part greater than q used by Wilcoxon greatly reduced the calculations required. For example, the values in the solution above of 1, 2, 3, 5, 7, 11, and 15 are obtained simply by finding the partitions of 1, 2, 3, 4, 5, 6, and 7, respectively. To illustrate how Wilcoxon simplified the calculations, consider $T = 35$ in the example above. What is necessary to compute is the number of unequal $q = 7$-part partitions of $T = 35$ with no part greater than $2q = (2)(7) = 14$. Since $r = 7$ corresponds to $T = 35$, as illustrated in Fig. 3.3, the number of unequal 7-part partitions of $T = 35$ is equivalent to the number of (equal or unequal) partitions of $r = 7$ with no part greater than $q = 7$. Table 3.3 lists the 15 unequal $q = 7$-part partitions of $T = 35$ with no part greater than $2q = 14$ and the corresponding 15 partitions of $r = 7$ with no part greater than $q = 7$.

3.10.2 Paired Samples

As with the unpaired data, Wilcoxon availed himself of a similar simplification for the case of paired data. Wilcoxon showed that for paired data, the number of unequal j-part partitions of r, with no part greater than i, was equal to the number of j-part partitions of $r - \binom{j}{2}$. For example, if $r = 10$, $j = 3$, and $i = 7$, then the unequal

Fig. 3.4 Analysis with $j = 1$ and $r - \binom{j}{2} = 8 - \binom{1}{2} = 8 - 0 = 8$

i	$i - 1 + 1$	List	$\mathbb{P}_1^i$
1	1	{1}	1
2	2	{2}	1
3	3	{3}	1
4	4	{4}	1
5	5	{5}	1
6	6	{6}	1
7	7	{7}	1
8	8	{8}	1
Sum			8

3-part partitions of $r = 10$ are $\{1, 2, 7\}$, $\{1, 3, 6\}$, $\{1, 4, 5\}$, and $\{2, 3, 5\}$, but the 3-part partitions of $10 - \binom{3}{2} = 10 - 3 = 7$ with no part greater than $i - j + 1 = 7 - 3 + 1 = 5$ are $\{1, 1, 5\}$, $\{1, 2, 4\}$, $\{1, 3, 3\}$, and $\{2, 2, 3\}$, which are much easier to enumerate and could readily be found in available tables of partitions.

Consider an example analysis on paired ranks with $q = 10$ paired differences and the sum of the negative differences between the ranks to be $T = -8$. The minimum value of T is zero when all the rank numbers are positive. The next possible sum is -1, when rank one receives a negative sign. As the sum of negative ranks increases, there are more and more ways in which a given total can be formed. The values for T and r are the same as both begin with zero. Then the one-sided probability value of $r = 8$ is given by

$$P = \left[1 + \sum_{j} \left(\sum_{i=j}^{r-\binom{j}{2}} \mathbb{P}_i^j \right) \right] \Big/ 2^q \ ,$$

where $\mathbb{P}_i^j$ represents the number of j-part partitions of i, q is the number of paired differences, and r is the serial number of the total under consideration in the series of possible totals. If in $\binom{j}{2}$ j is less than 2, $\binom{j}{2}$ is considered to be zero and if $\binom{j}{2}$ is greater than r, $r - \binom{j}{2}$ is undefined. For the example data with $r = 8$ and $q = 10$, Figs. 3.4, 3.5, and 3.6 illustrate the $\mathbb{P}_i^j$ for $j = 1, 2, 3$, respectively. Thus, for the example data, the equation is

$$P = \left[1 + \sum_{j} \left(\sum_{i=j}^{8-\binom{j}{2}} \mathbb{P}_i^j \right) \right] \Big/ 2^{10} \ ,$$

and the one-sided probability value is

$$P = \frac{1 + 8 + 12 + 4}{1{,}024} = \frac{25}{1{,}024} = 0.0244 \ ,$$

where the summations of the partitions yielding the sums of 8, 12, and 4 are illustrated in Figs. 3.4, 3.5, and 3.6.

Fig. 3.5 Analysis with $j = 2$ and $r - \binom{j}{2} = 8 - \binom{2}{2} = 8 - 1 = 7$

i	$i-2+1$	List	$\mathbb{P}_2^i$
1	0	——	0
2	1	$\{1,1\}$	1
3	2	$\{1,2\}$	1
4	3	$\{1,3\}\,\{2,2\}$	2
5	4	$\{1,4\}\,\{2,3\}$	2
6	5	$\{1,5\}\,\{2,4\}\,\{3,3\}$	3
7	6	$\{1,6\}\,\{2,5\}\,\{3,4\}$	3
Sum			12

Fig. 3.6 Analysis with $j = 3$ and $r - \binom{j}{2} = 8 - \binom{3}{2} = 8 - 3 = 5$

i	$i-3+1$	List	$\mathbb{P}_3^i$
1	-1	——	0
2	0	——	0
3	1	$\{1,1,1\}$	1
4	2	$\{1,1,2\}$	1
5	3	$\{1,1,3\}\,\{1,2,2\}$	2
Sum			4

3.11 Festinger and the Two-Sample Rank-Sum Test

The social psychologist, Leon Festinger, was also an accomplished statistician. In 1946 Festinger developed a new statistical test to evaluate differences between two independent means by first converting the data to ranks, a test that has largely been ignored [427]. This is unfortunate as, unlike the Wilcoxon test, Festinger's otherwise equivalent test allowed for unequal sample sizes.

L. Festinger

Leon Festinger is best known for his work in social psychology and, especially, his theories of cognitive dissonance and social comparisons, but Festinger was also a gifted statistician, working in the area of non-parametric statistics. Festinger was born in New York City and earned his B.Sc. degree in psychology from City College of New York in 1939, then moved to the University of Iowa to earn his Ph.D. in psychology in 1942.[31] After earning his Ph.D., Festinger worked first as a Research Associate at the University of Iowa, then joined the University of Rochester in 1943 as a Senior Statistician. In 1945, Festinger moved to the Massachusetts Institute of Technology,

(continued)

[31] Several sources list Festinger earning his Ph.D. in 1942 from Iowa State University, not the University of Iowa. Since his dissertation advisor was Kurt Lewin, who was at the University of Iowa from 1935 to 1944, the University of Iowa appears correct.

then to the University of Michigan in 1948, the University of Minnesota in 1951, Stanford University in 1955, and finally to the New School for Social Research (now, The New School) in 1968. Festinger remained at the New School until his death from liver cancer on 11 February 1989 at the age of 69 [1009, 1229].

In 1946 Festinger introduced a statistical test of differences between two independent means by first converting raw scores to ranks, then testing the difference between the means of the ranks [427]. Festinger provided tables for tests of significance based on exact probability values for the 0.05 and 0.01 confidence levels for $n = 2, \ldots, 15$, the smaller of the two samples, and $m = 2, \ldots, 38$, the larger sample. Festinger's approach to the two-sample rank-sum problem was developed independently of Wilcoxon's solution; moreover, Festinger's tables considered both equal and unequal sample sizes, whereas Wilcoxon's [1453] method allowed for only equal sample sizes. In addition, the approach that Festinger took was quite different from that of Wilcoxon. While both approaches generated all possible permutations of outcomes, Festinger's was considerably simpler to implement and is worth consideration here as a unique and ingenious recursive permutation generation method.

Consider two independent samples $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$ with $n \leq m$. Combining the samples x and y and assigning ranks to each case from 1 to $m + n$ structures the question as to the probability of obtaining any specified difference between sample ranks if both samples are drawn at random from the same population. Stated in terms of sums of ranks: what is the probability of obtaining any specified sum of ranks of n cases selected at random from the total of $m + n$ cases? The problem for Festinger was to generate exact probability distributions for sums of ranks given specified values of m and n.

For simplicity, consider first $m = 2$ and $n = 2$. The possible combinations of $m+n = 2+2 = 4$ considered $n = 2$ at a time are $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, and $\{3, 4\}$, yielding sums of 3, 4, 5, 5, 6, and 7, respectively. Thus, the frequency distribution of the sums is 3(1), 4(1), 5(2), 6(1), and 7(1), where the frequencies are enclosed in parentheses. If each case is independent of every other case and equally likely to be drawn, then each combination is equiprobable. However, as Festinger showed, there is an alternative way to generate this frequency distribution of sums. The frequency distribution of sums for $\binom{m+n}{n}$ can be constructed from the frequency distributions of sums for $\binom{m+n-1}{n}$ and $\binom{m+n-1}{n-1}$, as illustrated in Table 3.4.[32] The frequency distribution of $\binom{m+n-1}{n} = \binom{2+2-1}{2} = \binom{3}{2}$ is listed in Column 1 of

[32] The decomposition $\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$ has been well known since the publication of Blaise Pascal's *Traité du triangle arithmétique* in 1665, 3 years after his death [1088]. Thus, considering any one of n objects, $\binom{n-1}{r}$ gives the number of combinations that exclude it and $\binom{n-1}{r-1}$ the number of combinations that include it.

Table 3.4 Generation of frequency arrays for 3, 4, 5, 6, and 7 objects considered $n = 2$ at a time

	Column								
	1	2	3	4	5	6	7	8	9
Sum	$\binom{3}{2}$	$\binom{3}{1}$	$\binom{4}{2}$	$\binom{4}{1}$	$\binom{5}{2}$	$\binom{5}{1}$	$\binom{6}{2}$	$\binom{6}{1}$	$\binom{7}{2}$
3	1		1		1		1		1
4	1		1		1		1		1
5	1	1	2		2		2		2
6		1	1	1	2		2		2
7		1	1	1	2	1	3		3
8				1	1	1	2	1	3
9				1	1	1	2	1	3
10						1	1	1	2
11						1	1	1	2
12								1	1
13								1	1

Table 3.4 and the frequency distribution of sums for $\binom{m+n-1}{n-1} = \binom{2+2-1}{2-1} = \binom{3}{1}$ is listed in Column 2 of Table 3.4. Note that the frequency distribution of sums for $\binom{3}{1}$ is offset from the frequency distribution of sums for $\binom{3}{2}$. Since the sum of ranks below the value 5 would not be affected by the addition of a 4th case to the ranks of $\binom{3}{2}$, only the totals of 5, 6, and 7 would be augmented by one or more possibilities. In general, the starting value for frequency distribution $\binom{m+n-1}{n-1}$ is given by $n(n+1)/2 + m$; in this case, $2(2+1)/2 + 2 = 5$. Thus, the frequency distribution of sums for $\binom{m+n}{n} = \binom{4}{2}$ in Column 3 is constructed from the frequency distributions of sums for $\binom{m+n-1}{n} = \binom{3}{2}$ and $\binom{m+n-1}{n-1} = \binom{3}{1}$ in Columns 1 and 2 in Table 3.4, respectively, by simply adding across Columns 1 and 2 to obtain the frequency distribution of sums for $\binom{4}{2}$ in Column 3.

Now consider the frequency distribution of sums for $m = 3$ and $n = 2$. The possible combinations of $m + n = 5$ considered $n = 2$ at a time are $\{1,2\}$, $\{1,3\}$, $\{1,4\}$, $\{1,5\}$, $\{2,3\}$, $\{2,4\}$, $\{2,5\}$, $\{3,4\}$, $\{3,5\}$, and $\{4,5\}$, yielding sums of 3, 4, 5, 6, 5, 6, 7, 7, 8, and 9, respectively. The frequency distribution of the sums is therefore 3(1), 4(1), 5(2), 6(2), 7(2), 8(1), and 9(1). The frequency distribution of sums for $\binom{m+n}{n} = \binom{3+2}{2} = \binom{5}{2}$ in Column 5 of Table 3.4 can be constructed from the frequency distributions of sums for $\binom{m+n-1}{n} = \binom{3+2-1}{2} = \binom{4}{2}$ and $\binom{m+n-1}{n-1} = \binom{3+2-1}{2-1} = \binom{4}{1}$ in Columns 3 and 4, respectively, in Table 3.4. In similar fashion to the previous case, no sum of ranks below the value 6 would be affected by the addition of a 5th case to the sum of ranks for $\binom{4}{2}$, thus the starting position for the frequency distribution of $\binom{4}{1}$ in Column 4 is the value $n(n+1)/2 + m = 2(2+1)/2 + 3 = 6$. Again, the frequency distribution of sums for $\binom{m+n}{n} = \binom{5}{2}$ in Column 5 is constructed from the frequency distributions of sums for $\binom{m+n-1}{n} = \binom{4}{2}$ and $\binom{m+n-1}{n-1} = \binom{4}{1}$ in Columns 3 and 4 in Table 3.4, respectively, by adding across Columns 3 and 4 to obtain the frequency distribution of sums for $\binom{5}{2}$ in Column 5.

Table 3.5 Generation of frequency arrays for 4, 5, 6, 7, and 8 objects considered $n = 3$ at a time

	Column								
	1	2	3	4	5	6	7	8	9
Sum	$\binom{4}{3}$	$\binom{4}{2}$	$\binom{5}{3}$	$\binom{5}{2}$	$\binom{6}{3}$	$\binom{6}{2}$	$\binom{7}{3}$	$\binom{7}{2}$	$\binom{8}{3}$
6	1		1		1		1		1
7	1		1		1		1		1
8	1	1	2		2		2		2
9	1	1	2	1	3		3		3
10		2	2	1	3	1	4		4
11		1	1	2	3	1	4	1	5
12		1	1	2	3	2	5	1	6
13				2	2	2	4	2	6
14				1	1	3	4	2	6
15				1	1	2	3	3	6
16						2	2	3	5
17						1	1	3	4
18						1	1	2	3
19								2	2
20								1	1
21								1	1

In this manner, Festinger was able to recursively generate all frequency distributions of sums for $m + n$ objects considered $n = 2$ at a time. In addition to the frequency distributions of sums for $\binom{4}{2}$ and $\binom{5}{2}$, Table 3.4 illustrates the construction of the frequency distribution of sums for $\binom{6}{2}$ in Column 7 from the frequency distributions of sums for $\binom{5}{2}$ and $\binom{5}{1}$ in Columns 5 and 6, respectively, and the frequency distribution of sums for $\binom{7}{2}$ in Column 9 from the frequency distributions of sums for $\binom{6}{2}$ and $\binom{6}{1}$ in Columns 7 and 8, respectively. Thus, for example, with $m = 4$, $m > n = 2$, and $m + n = 4 + 2 = 6$ the sum of 7 can occur in only three ways: $\{1, 6\}$, $\{2, 5\}$, and $\{3, 4\}$. As illustrated in Table 3.4, the frequency 3 is read in Column 7 with heading $\binom{6}{2}$ in the row designated as Sum 7. The probability, therefore, of a sum of 7 is $3/\binom{m+n}{n} = 3/\binom{4+2}{2} = 3/\binom{6}{2} = 3/15 = 0.20$.

Once the exact frequency distributions of sums for $m + n$ ranks considered $n = 2$ at a time are established, it is relatively straightforward to construct exact frequency distributions of sums for $m + n$ ranks considered $n = 3$ at a time, using the same method. Table 3.5 illustrates the construction of the frequency distribution of sums for $\binom{m+n}{n} = \binom{3+2}{3} = \binom{5}{3}$ in Column 3 from the frequency distributions of sums for $\binom{m+n-1}{n} = \binom{4}{3}$ and $\binom{m+n-1}{n-1} = \binom{4}{2}$ in Columns 1 and 2, respectively. In like manner, the frequency distribution of sums for $\binom{6}{3}$ in Column 5 is constructed from the frequency distributions of sums for $\binom{5}{3}$ and $\binom{5}{2}$ in Columns 3 and 4, respectively; the frequency distribution of sums for $\binom{7}{3}$ in Column 7 is constructed from the frequency distributions of sums for $\binom{6}{3}$ and $\binom{6}{2}$ in Columns 5 and 6, respectively; and the frequency distribution of sums for $\binom{8}{3}$ in Column 9 is constructed from the frequency distributions of sums for $\binom{7}{3}$ and $\binom{7}{2}$ in Columns 7 and 8, respectively. As before, the frequency distribution of sums for $\binom{m+n-1}{n-1}$ is offset and has a starting value given by $n(n+1)/2+m$, e.g., for $\binom{8}{3}$ the starting value for $\binom{7}{2}$ is $3(3+1)/2+5=11$.

This method allowed Festinger to recursively generate exact frequency distributions of sums for any combination of $m+n$ and n. For example, to obtain the exact frequency distribution of the sum of $n = 7$ cases selected at random from $m+n = 18$ ranked cases with $n = 7$ and $m = 11$, add to the distribution of the sums of $n = 7$ cases from $m+n-1 = 17$ ranked cases, the distribution of the sums of $n-1 = 6$ cases from the $m+n-1 = 17$ ranked cases, making the first addition for the sum equal to $n(n+1)/2+m = 7(7+1)/2+11 = 39$, which is the lowest sum where the frequency sums will be affected. Festinger explained that since the distributions of sums were symmetrical about

$$\frac{n(m+n+1)}{2} ,$$

only one-half of the distribution need be computed.

Finally, Festinger proposed a convenient alternative for summarizing and presenting the frequency distributions of sums. He replaced the sums of ranks of the smaller of the two samples with the absolute deviation (d) of the mean of the ranks of the smaller sample from the mean of the ranks of the total group, using

$$d = \left| \frac{1}{n} \sum_{i=1}^{n} R_i - \frac{m+n+1}{2} \right| , \tag{3.3}$$

where n is the number of cases in the smaller sample, $m+n$ is the number of cases in both samples combined, and $\sum_{i=1}^{n} R_i$ is the sum of the ranks of the cases in the smaller sample. The last term in Eq. (3.3) is, of course, the mean of the $m+n$ ranks. Festinger then presented two tables containing the d values necessary for tests of significance at the 0.01 and 0.05 levels of confidence. For values of n from 2 to 12, the Festinger tables listed values of d from $m = 2$ to $m = 38$ [427].

3.12 Mann–Whitney and a Two-Sample Rank-Sum Test

Henry Mann and his graduate student, Donald Whitney, published a two-sample rank-sum test in 1947 that was equivalent to the two-sample rank-sum test proposed by Wilcoxon 2 years prior, but was easier to calculate, allowed for unequal sample sizes, and also permitted larger samples than the Wilcoxon two-sample rank-sum test [880].

D.R. Whitney

While Henry Mann (q.v. page 125) was at The Ohio State University from 1946 to 1964, one of his graduate students was Donald Ransom Whitney. Whitney had earned his B.A. degree in mathematics from Oberlin College in 1936 and his M.S. degree in mathematics from Princeton University in 1939. After service in the Navy during World War II, Whitney enrolled in the Ph.D. program at The Ohio State University in 1946, where eventually he came to work under Henry Mann. After receiving his Ph.D. in mathematics in 1949, Whitney remained at The Ohio State University, eventually becoming Chair of the newly established Department of Statistics in 1974. Whitney retired from The Ohio State University in 1982, whereupon he received the University Distinguished Service Award. Donald Ransom Whitney passed away on 16 August 2007 at the age of 92 [1460].

In 1947 Mann and Whitney, acknowledging the previous work by Wilcoxon on the two-sample rank-sum test [1453], proposed an equivalent test statistic, U, based on the relative ranks of two samples denoted by $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ [880].[33] Like Festinger in 1946, Mann and Whitney utilized a recurrence relation involving n and m and, using this relation, computed tables of exact probability values for U up to $n = m = 8$, many more, they noted, than the few probability values provided by Wilcoxon. As Mann and Whitney explained, let the measurements $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ be arranged in order and let U count the number of times a y precedes an x. For example, given $n = 4$ x values and $m = 2$ y values, consider the sequence $\{x, y, x, x, y, x\}$ where $U = 4$: the first y precedes three x values and the second y precedes one x value; thus, $U = 3 + 1 = 4$. Also, let the Wilcoxon statistic, W, be the sum of the m rank-order statistics $\{y_1, y_2, \ldots, y_m\}$. The relationship between Wilcoxon's W statistic and Mann and Whitney's U statistic can be expressed as

$$U = mn + \frac{m(m+1)}{2} - W \ ,$$

where $0 \leq U \leq mn$. Mann and Whitney noted that since Wilcoxon only considered the case of $n = m$, it seemed worthwhile to extend this important work to $n \neq m$ and larger values of n and m.

Consider again the ordered sequences of n x and m y values, replace each x with a 0 and each y with a 1, let U denote the number of times a 1 precedes a 0, and let $\bar{p}_{n,m}(U)$ represent the number of sequences of n 0s and m 1s in each of which a

[33] A particularly clear exposition of the Mann–Whitney U test is given in a 1952 paper by Lincoln Moses on "Non-parametric statistics for psychological research" published in *Psychological Bulletin* [1010].

Table 3.6 Sequences of $n = 4$ 0s and $m = 2$ 1s for $\bar{p}_{n,m}(U)$, $\bar{p}_{n-1,m}(U-m)$, and $\bar{p}_{n,m-1}(U)$

	$\bar{p}_{n,m}(U)$		$\bar{p}_{n-1,m}(U-m)$		$\bar{p}_{n,m-1}(U)$	
Row	Sequence	U	Sequence	U	Sequence	U
1	0 0 0 0 1 1	0	0 0 0 1 1	0	0 0 0 0 1	0
2	0 0 0 1 0 1	1	0 0 1 0 1	1	0 0 0 1 0	1
3	0 0 1 0 0 1	2	0 1 0 0 1	2	0 0 1 0 0	2
4	0 1 0 0 0 1	3	1 0 0 0 1	3	0 1 0 0 0	3
5	1 0 0 0 0 1	4	0 0 1 1 0	2	1 0 0 0 0	4
6	0 0 0 1 1 0	2	0 1 0 1 0	3		
7	0 0 1 0 1 0	3	1 0 0 1 0	4		
8	0 1 0 0 1 0	4	0 1 1 0 0	4		
9	1 0 0 0 1 0	5	1 0 1 0 0	5		
10	0 0 1 1 0 0	4	1 1 0 0 0	6		
11	0 1 0 1 0 0	5				
12	1 0 0 1 0 0	6				
13	0 1 1 0 0 0	6				
14	1 0 1 0 0 0	7				
15	1 1 0 0 0 0	8				

1 precedes a 0 U times. For example, suppose the sequence is $\{1, 1, 0, 0, 1, 0\}$, then $U = 7$ as the first 1 precedes three 0 values, the second 1 precedes the same three 0 values, and the third 1 precedes only one 0 value. Mann and Whitney then developed the recurrence relation,

$$\bar{p}_{n,m}(U) = \bar{p}_{n-1,m}(U-m) + \bar{p}_{n,m-1}(U) \ , \tag{3.4}$$

where $\bar{p}_{n-1,m}(U-m) = 0$ if $U \leq m$.

An example of the recurrence relation will illustrate the Mann–Whitney procedure. Table 3.6 lists all the sequences of 0s and 1s and corresponding values of U for $\bar{p}_{n,m}(U)$, $\bar{p}_{n-1,m}(U-m)$, and $\bar{p}_{n,m-1}(U)$ for $n = 4$ and $m = 2$. There are $\binom{m+n}{m} = \binom{2+4}{2} = 15$ values of U in the first sequence of 0s and 1s in Table 3.6, $\binom{m+n-1}{m} = \binom{2+4-1}{2} = 10$ values of U in the second sequence of 0s and 1s, and $\binom{m-1+n}{m-1} = \binom{2-1+4}{2-1} = 5$ values of U in the third sequence of 0s and 1s.[34] To illustrate the recurrence process with $U = 3$, $\bar{p}_{n,m}(3) = 2$, as there are two occurrences of $U = 3$ (in Rows 4 and 7) in the leftmost column of sequences in Table 3.6. Then, $\bar{p}_{n-1,m}(U-m) = \bar{p}_{4-1,2}(3-2) = 1$, as there is only a single occurrence of $U = 1$ (in Row 2) in the middle column of sequences in Table 3.6, and $\bar{p}_{n,m-1}(U) = \bar{p}_{4,2-1}(3) = 1$, as there is only a single occurrence of $U = 3$ (in Row 4) in the rightmost column of sequences in Table 3.6. Then, following Eq. (3.4), $2 = 1 + 1$.

[34] Here, the decomposition is identical to Festinger's, as given in [427].

Given that under the null hypothesis each of the $(n+m)!/(n!\, m!)$ sequences of n 0s and m 1s is equally-likely, let $p_{n,m}(U)$ represent the probability of a sequence in which a 1 precedes a 0 U times. For example, for $U = 3$ in the leftmost column of sequences of 0s and 1s in Table 3.6,

$$p_{n,m}(U) \times \frac{n!\, m!}{(n+m)!} = p_{4,2}(3) \times \frac{4!\, 2!}{(4+2)!} = \frac{2}{15} = 0.1333 \,.$$

Mann and Whitney also provided a recurrence relation for the probability values of U given by

$$p_{n,m}(U) = \frac{n}{n+m} p_{n-1,m}(U-m) + \frac{m}{n+m} p_{n,m-1}(U) \,,$$

where

$$p_{n-1,m}(U-m) = \bar{p}_{n-1,m}(U-m) \times \frac{(n-1)!\, m!}{(n+m-1)!}$$

and

$$p_{n,m-1}(U) = \bar{p}_{n,m-1}(U) \times \frac{n!\, (m-1)!}{(n+m-1)!} \,.$$

Thus, for $U = 3$ in Table 3.6,

$$\begin{aligned} p_{4,2}(3) &= \frac{4}{4+2}\, p_{4-1,2}(3-2) + \frac{2}{4+2}\, p_{4,2-1}(3) \\ \frac{2}{15} &= \left(\frac{4}{6}\right)\left(\frac{1}{10}\right) + \left(\frac{2}{6}\right)\left(\frac{1}{5}\right) \\ \frac{2}{15} &= \frac{1}{15} + \frac{1}{15} \,. \end{aligned}$$

Mann and Whitney used this recurrence relation to construct tables of exact probability values up to and including $n = m = 8$. Finally, from the recurrence relation Mann and Whitney derived explicit expressions for the mean, variance, and various higher moments for U, and explained that the limit of the distribution is normal if $\min(n, m) \to \infty$ [880].

It should be noted that in 1914 Gustav Deuchler suggested an approach that was essentially the same as that used by Mann and Whitney in their treatment of the two-sample rank-sum test [345]. Deuchler's work in this area seems to have been neglected, but William Kruskal attempted to redress this failure in a 1957 article on "Historical notes on the Wilcoxon unpaired two-sample test" in *Journal of the American Statistical Association* [776]. In a 1952 article W.H. Kruskal and W.A. Wallis provided a list of independent discoveries of the Wilcoxon two-sample rank-sum test [779] and this 1957 article is, in part, an attempt to update that list.

Fig. 3.7 Rankings of a dichotomous variable

1	2	3	4	5	6
$-$	$+$	$-$	$-$	$-$	$+$

Also mentioned in the 1957 article, but omitted in the 1952 article, was a 1947 article by J.W. Whitfield who independently discovered the Mann–Whitney test [1443].

3.13 Whitfield and a Measure of Ranked Correlation

In 1947 John W. Whitfield proposed a measure of rank-order correlation between two variables, one of which was composed of ranks and the other dichotomous [1443].[35] While not presented as a permutation test per se, the article by Whitfield is of historical importance as it is occasionally cited as an independent discovery of the Wilcoxon two-sample rank-sum test [e.g., 776, pp. 358–359]. Whitfield considered the dichotomous variable as a ranking composed entirely of two sets of tied rankings. An example will illustrate the procedure. Following Whitfield, consider the rank data in Fig. 3.7 where the $-$ and $+$ signs indicate the dichotomous variable and the ranks are from 1 to 6. Let $m = 2$ denote the number of ranks in the "$+$" group and let $n = 4$ denote the number of ranks in the "$-$" group.

Now consider the $n = 4$ ranks in the group identified by a $-$ sign: 1, 3, 4, and 5. Beginning with rank 1 with a $-$ sign, there are no ranks with a $+$ sign to the left of rank 1 and two ranks with a $+$ sign to the right of rank 1 (ranks 2 and 6); so compute $0-2 = -2$. For rank 3 with a $-$ sign, there is one rank to the left of rank 3 with a $+$ sign (rank 2) and one rank to the right of rank 3 with a $+$ sign (rank 6); so compute $1 - 1 = 0$. For rank 4 with a $-$ sign, there is one rank to the left of rank 4 with a $+$ sign (rank 2) and one rank to the right of rank 4 with a $+$ sign (rank 6); so compute $1 - 1 = 0$. Finally, for rank 5 with a $-$ sign, there is one rank to the left of rank 5 with a $+$ sign (rank 2) and one rank to the right of rank 5 with a $+$ sign (rank 6); so compute $1 - 1 = 0$. The sum of the differences is $S = -2 + 0 + 0 + 0 = -2$. In this manner, Whitfield's approach incorporated unequal sample sizes with $m \neq n$ as well as tied ranks.

Since the number of possible pairs of $m + n$ consecutive integers is given by $(m + n)(m + n - 1)/2$, Whitfield defined and calculated his test statistic as

$$\tau = \frac{2S}{(m+n)(m+n-1)} = \frac{2(-2)}{(2+4)(2+4-1)} = \frac{-4}{30} = -0.1333\ .$$

[35] Whitfield's article was followed immediately in the same issue of *Biometrika* with a comment by M.G. Kendall noting that "Mr Whitfield has correctly surmised the variance [of τ] when one ranking contains ties, and the other is a dichotomy" [733, p. 297].

Table 3.7 Fifteen paired observations with concordant/discordant (C/D) pairs and associated pair values

Number	Pair	C/D	Value	Number	Pair	C/D	Value
1	1–2	$-,+$	-1	9	2–6	$+,+$	0
2	1–3	$-,-$	0	10	3–4	$-,-$	0
3	1–4	$-,-$	0	11	3–5	$-,-$	0
4	1–5	$-,-$	0	12	3–6	$-,+$	-1
5	1–6	$-,+$	-1	13	4–5	$-,-$	0
6	2–3	$+,-$	$+1$	14	4–6	$-,+$	-1
7	2–4	$+,-$	$+1$	15	5–6	$-,+$	-1
8	2–5	$+,-$	$+1$				

Whitfield's S is directly related to the U statistic of Mann and Whitney [880] and, hence, to the W statistic of Wilcoxon [1453].[36] Compare statistic S with the U statistic of Mann and Whitney. For the data in Fig. 3.7 there are $m = 2$ + signs and $n = 4$ − signs, so considering the lesser of the two (the $m = 2$ + signs), the first + sign (rank 2) precedes three − signs (ranks 3–5) and the second + sign precedes no − signs, so $U = 3 + 0 = 3$. The relationship between Whitfield's S and Mann and Whitney's U is given by $S = 2U - mn$ [229, 776]; thus, $S = 2(3) - (2)(4) = 6 - 8 = -2$. For the example data in Fig. 3.7, the Wilcoxon's W test statistic for the smaller of the two sums (with the $m = 2$ + signs) is $W = 2 + 6 = 8$ and the relationship with S is given by $S = m(m + n + 1) - 2W$; thus, $S = 2(2 + 4 + 1) - (2)(8) = 14 - 16 = -2$.

As Whitfield mentioned, the calculation of S was fashioned after a procedure first introduced by Kendall in 1945[37] and Whitfield was apparently unaware of the two-sample rank-sum tests published by Wilcoxon in 1945, Festinger in 1946, and Mann and Whitney in 1947, as they are not referenced in the Whitfield article. Kendall considered the number of concordant (C) and discordant (D) pairs, of which there is a total of $(m + n)(m + n - 1)/2$ pairs when there are no ties in the $m + n$ consecutive integers [730]. For the example data in Fig. 3.7 there are $(2+4)(2+4-1)/2 = 15$ pairs. Table 3.7 numbers and lists the 15 pairs, the concordant/discordant classification of pairs, and the pair values, where concordant pairs ($-,-$ and $+,+$) are given a value of 0, and discordant pairs ($+,-$ and $-,+$) are given values of $+1$ and -1, respectively. The sum of the pair values in Table 3.7 for the 15 pairs is $S = -5 + 3 = -2$.

Today it is well-known, although poorly documented, that when one classification is a dichotomy and the other classification is ordered, with or without tied values, the S statistic of Kendall is equivalent to the Mann–Whitney U statistic; see also articles by Lincoln Moses in 1956 and Edmund John Burr in 1960 on this topic

[36] In 1968 Charles R. Kraft and Constance van Eeden showed how Kendall's τ can be computed as a sum of Wilcoxon W statistics [768, pp. 180–181].

[37] Whitfield lists the date of the Kendall article as 1946, but Kendall's article was actually published in *Biometrika* in 1945.

[229, 1011]. Whitfield was apparently the first to discover the relationship between S, the statistic underlying Kendall's τ rank-order correlation coefficient, and U, the Mann–Whitney two-sample rank-sum statistic for two independent samples. However, it was Hemelrijk in 1952 [610] and Jonckheere in 1954 [699] who made the relationship explicit; see also a discussion by Leach in 1979 [806, p. 183]. Because the Jonckheere–Terpstra test, when restricted to two independent samples, is mathematically identical in reverse application to the Wilcoxon and Mann–Whitney tests (see [699, p. 138] and [1153, p. 396]), the two-sample rank-sum test is sometimes referred to as the Kendall–Wilcoxon–Mann–Whitney–Jonckheere–Festinger test [1011, p. 246]. Whitfield concluded his article with derivations of the variances of S for both untied and tied rankings and included a correction for continuity. For untied ranks the variance of S, as given by Kendall [731], is

$$\sigma_S^2 = \frac{(m+n)(m+n-1)[2(m+n)+5]}{18}$$

and the desired probability value is obtained from the asymptotically $N(0,1)$ distribution when $\min(m,n) \to \infty$. For the example data listed in Fig. 3.7, the variance of S is calculated as

$$\sigma_S^2 = \frac{(2+4)(2+4-1)[2(2+4)+5]}{18} = 28.3333$$

and

$$\tau = \frac{S}{\sigma_S} = \frac{-2}{\sqrt{28.3333}} = -0.3757 \ ,$$

with a one-sided probability value of 0.3536.

3.13.1 An Example of Whitfield's Approach

It is common today to transform a Pearson correlation coefficient between two variables (r_{xy}) into Student's pooled t test for two independent samples and vice-versa, i.e.,

$$t = r_{xy}\sqrt{\frac{m+n-2}{1-r_{xy}^2}} \quad \text{and} \quad r_{xy} = \frac{t}{\sqrt{t^2+m+n-2}} \ ,$$

where m and n indicate the number of observations in Samples 1 and 2, respectively. It appears that Whitfield was the first to transform Kendall's rank-order correlation coefficient, τ, into Mann and Whitney's two-sample rank-sum test, U, for two independent samples. Actually, since

$$\tau = \frac{2S}{(m+n)(m+n-1)} \ ,$$

Age:	20	20	20	20	22	23	23	24	25	25	25	25	27	29	29	35	35
Rank:	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	5	$6\frac{1}{2}$	$6\frac{1}{2}$	9	$10\frac{1}{2}$	$10\frac{1}{2}$	$10\frac{1}{2}$	$10\frac{1}{2}$	13	$14\frac{1}{2}$	$14\frac{1}{2}$	$16\frac{1}{2}$	$16\frac{1}{2}$
Sample:	*A*	*A*	*A*	*A*	*B*	*A*	*A*	*B*	*A*	*A*	*A*	*A*	*B*	*A*	*A*	*B*	*B*

Fig. 3.8 Listing of the $m + n = 17$ age and rank scores from Samples *A* and *B*

A	4	0	2	0	4	0	2	0	12
B	0	1	0	1	0	1	0	2	5
	4	1	2	1	4	1	2	2	17

Fig. 3.9 Contingency table of the frequency of ranks in Fig. 3.8

Whitfield established the relationship between the variable part of Kendall's τ, S, and Mann and Whitney's U. To show just how Whitfield accomplished this, consider the data listed in Fig. 3.8. The data consist of $m = 12$ adult ages from Sample A and $n = 5$ adult ages from Sample B, with associated ranks. The sample membership of the ages/ranks is indicated by an A or a B immediately beneath the rank score.

Now, arrange the two samples into a contingency table with two rows and columns equal to the frequency distribution of the combined samples, as in Fig. 3.9. Here the first row of frequencies in Fig. 3.9 represents the runs in the list of ranks in Fig. 3.8 labeled as A, i.e., there are four values of $2\frac{1}{2}$, no value of 5, two values of $6\frac{1}{2}$, no value of 9, four values of $10\frac{1}{2}$, and so on. The second row of frequencies in Fig. 3.9 represents the runs in the list of ranks in Fig. 3.8 labeled as B, i.e., there is no value of $2\frac{1}{2}$ labeled as B, one value of 5, no value of $6\frac{1}{2}$, one value of 9, and so on. Finally, the column marginal totals are simply the sums of the two rows. This contingency arrangement permitted Whitfield to transform a problem of the difference between two independent samples into a problem of correlation between two sets of ranks.

Denote by $\mathbf{X}$ the $r \times c$ table in Fig. 3.9 with $r = 2$ and $c = 8$ and let x_{ij} indicate a cell frequency for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. Then, as noted by E.J. Burr in 1960, S can be expressed as the algebraic sum of all second-order determinants in $\mathbf{X}$ [229]:

$$S = \sum_{i=1}^{r-1} \sum_{j=i+1}^{r} \sum_{k=1}^{c-1} \sum_{l=k+1}^{c} \left(x_{ik}x_{jl} - x_{il}x_{jk} \right) .$$

Thus, for the data listed in Fig. 3.9 there are $c(c-1)/2 = 8(8-1)/2 = 28$ second-order determinants:

$$S = \begin{vmatrix} 4 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} 4 & 2 \\ 0 & 0 \end{vmatrix} + \begin{vmatrix} 4 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} 4 & 4 \\ 0 & 0 \end{vmatrix} + \begin{vmatrix} 4 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} 4 & 2 \\ 0 & 0 \end{vmatrix} + \cdots + \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} .$$

Therefore,

$$\begin{aligned} S = {} & (4)(1) - (0)(0) + (4)(0) - (2)(0) + (4)(1) - (0)(0) + (4)(0) - (4)(0) \\ & + (4)(1) - (0)(0) + (4)(0) - (2)(0) + (4)(2) - (0)(0) + (0)(0) - (2)(1) \\ & + (0)(1) - (0)(1) + (0)(0) - (4)(1) + (0)(1) - (0)(1) + (0)(0) - (2)(1) \\ & + (0)(2) - (0)(1) + (2)(1) - (0)(0) + (2)(0) - (4)(0) + (2)(1) - (0)(0) \\ & + (2)(0) - (2)(0) + (2)(2) - (0)(0) + (0)(0) - (4)(1) + (0)(1) - (0)(1) \\ & + (0)(0) - (2)(1) + (0)(2) - (0)(1) + (4)(1) - (0)(0) + (4)(0) - (2)(0) \\ & + (4)(2) - (0)(0) + (0)(0) - (2)(1) + (0)(2) - (0)(1) + (2)(2) - (0)(0) \end{aligned}$$

and $S = 4 + 0 + 4 + \cdots + 2 + 4 = 28$.

Alternatively, as Kendall showed in 1948 [734], the number of concordant pairs is given by

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} x_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} x_{kl} \right)$$

and the number of discordant pairs is given by

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} x_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} x_{kl} \right) .$$

Thus, for **X** in Fig. 3.9, C is calculated by proceeding from the upper-left cell with frequency $x_{11} = 4$ downward and to the right, multiplying each cell frequency by the sum of all cell frequencies below and to the right, and summing the products, i.e.,

$$\begin{aligned} C = {} & (4)(1 + 0 + 1 + 0 + 1 + 0 + 2) + (0)(0 + 1 + 0 + 1 + 0 + 2) \\ & + (2)(1 + 0 + 1 + 0 + 2) + (0)(0 + 1 + 0 + 2) \\ & + (4)(1 + 0 + 2) + (0)(0 + 2) + (2)(2) \\ & = 20 + 0 + 8 + 0 + 12 + 0 + 4 = 44 , \end{aligned}$$

and D is calculated by proceeding from the upper-right cell with frequency $x_{19} = 0$ downward and to the left, multiplying each cell frequency by the sum of all cell frequencies below and to the left, and summing the products, i.e.,

$$\begin{aligned} D = {} & (0)(0 + 1 + 0 + 1 + 0 + 1 + 0) + (2)(1 + 0 + 1 + 0 + 1 + 0) \\ & + (0)(0 + 1 + 0 + 1 + 0) + (4)(1 + 0 + 1 + 0) \\ & + (0)(0 + 1 + 0) + (2)(1 + 0) + (0)(0) \\ & = 0 + 6 + 0 + 8 + 0 + 2 + 0 = 16 . \end{aligned}$$

Then, as defined by Kendall, $S = C - D = 44 - 16 = 28$.

To calculate Mann and Whitney's U for the data listed in Fig. 3.8, the number of A ranks to the left of (less than) the first B is 4; the number of A ranks to the left of the second B is 6; the number of A ranks to the left of the third B is 10; and the number of A ranks to the left of the fourth and fifth B are 12 each. Then $U = 4+6+10+12+12 = 44$. Finally, $S = 2U - mn = (2)(44) - (12)(5) = 28$. Thus, Kendall's S statistic, as redefined by Whitfield, includes as special cases Yule's Q test for association in 2×2 contingency tables and the Mann–Whitney two-sample rank-sum U test for larger $r \times c$ contingency tables.

It is perhaps not surprising that Whitfield established a relationship between Kendall's S and Mann and Whitney's U as Mann published a test for trend in 1945 (q.v. page 125) that was identical to Kendall's S, as Mann noted [879]. The Mann test is known today as the Mann–Kendall test for trend where for n values in an ordered time series $x_1, \ldots, x_n$,

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}\left(x_i - x_j\right) ,$$

where

$$\operatorname{sgn}(\cdot) = \begin{cases} +1 & \text{if } x_i - x_j > 0 , \\ 0 & \text{if } x_i - x_j = 0 , \\ -1 & \text{if } x_i - x_j < 0 . \end{cases}$$

3.14 Olmstead–Tukey and the Quadrant-Sum Test

In 1947 Paul Olmstead and John Tukey (q.v. page 232) proposed a new test for the association of two continuous variables [1059].[38] They termed the new test the "quadrant-sum test," but it is better known as the "corner test for association." Olmstead and Tukey observed that when a moderate number of paired observations (25–200) on two quantities were plotted as a scatter diagram, visual examination tended to give greater weight to observations near the periphery of the scatter diagram. They pointed out that a quantitative test of association with such concentration on the periphery was lacking, and the quadrant-sum test was developed to fill this gap [1059, p. 499]. In classic Tukey fashion, they recommended the quadrant-sum test for exploratory investigations of large data sets, due to its simplicity and ease of use; see also discussions by Tukey and Olmstead in 1947 [1383], Mood in 1950 [1000, pp. 410–414], and Daniel in 1978 [313, pp. 321–324].

[38] It was common at this time to assume continuous variables as this ensured no tied values, cf. articles by Sun and Sherman in 1996 [1335, p. 90] and Gebhard and Schmitz in 1998 [502, p. 76].

Consider a sample of n paired x and y values given by

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) .$$

Plot the sample values in an xy plane and draw a horizontal line at $x = x_m$ and a vertical line at $y = y_m$, where x_m (y_m) is the median of the x (y) values without regard to the values of y (x). Label the quadrants as $+$, $-$, $+$, $-$, beginning in the upper right-hand quadrant and moving counterclockwise, so that the upper right and lower left quadrants are positive. Then beginning at the right side of the scatter diagram with the sample point furthest from the vertical line, count in (in order of abscissae) along the observations until forced to cross y_m. Write down the number of observations encountered before crossing y_m, attaching a $+$ sign if the observation lies in the $+$ quadrant and a $-$ sign if the observation lies in the $-$ quadrant. Denote the count by s_1. Do a similar count, moving from the top sample point downward, another count moving from the leftmost sample point to the right, and a final count moving from the bottom sample point upward. Let the number of points be denoted by s_2, s_3, and s_4, respectively, with attached $+$ or $-$ signs depending on whether the sample points in each case fall into a $+$ or $-$ quadrant. Finally, let S denote the algebraic (quadrant) sum of s_1, s_2, s_3, and s_4, with their respective signs attached. Note that the order of $s_1, \ldots, s_4$ is not important and also that an observation may be counted twice, once in counting from the top, say, and again when counting from the right.

Olmstead and Tukey explained that the set of x values, the set of y values, and the permutations of the order of the y values when the pairs were ordered by the x values were independently distributed, and that any permutation was as likely as any other permutation. Thus, since the quadrant sum S depended only on the permutation, its distribution in the absence of association did not depend on the distribution of x and y. The question for Olmstead and Tukey was: how many permutations yield a count of exactly k positive values? They tabulated exact probability values of $P(|S| \geq k)$ for $n = 2, 3, 4, 5, 7$ and for $k = 1, 2, \ldots, 30$ and showed that for large n

$$\lim_{n \to \infty} P(|S| \geq k) = \frac{9k^3 + 9k^2 + 168k + 208}{(216)(2^k)} .$$

Olmstead and Tukey also provided extensions to higher dimensions and applications to serial correlation.

Consider an example with 28 paired observations as depicted in Fig. 3.10. Beginning at the right side of the diagram, count in along the observations, moving toward the center until forced to cross the horizontal median (y_m) and write down the number of observations met before the crossing the median (the dashed line), attaching a $+$ $(-)$ sign if the observations lie in the $+$ $(-)$ quadrant. In this example, $s_1 = +2$, as the observations are in the $+$ quadrant. Then, moving from the bottom toward the center, $s_2 = +1$, as the observation is in the $+$ quadrant. Moving from the left side, $s_3 = +5$ and moving from the top, $s_4 = +3$. Thus, the quadrant sum is $S = 2 + 1 + 5 + 3 = 11$, yielding an approximate probability value of 0.0342.

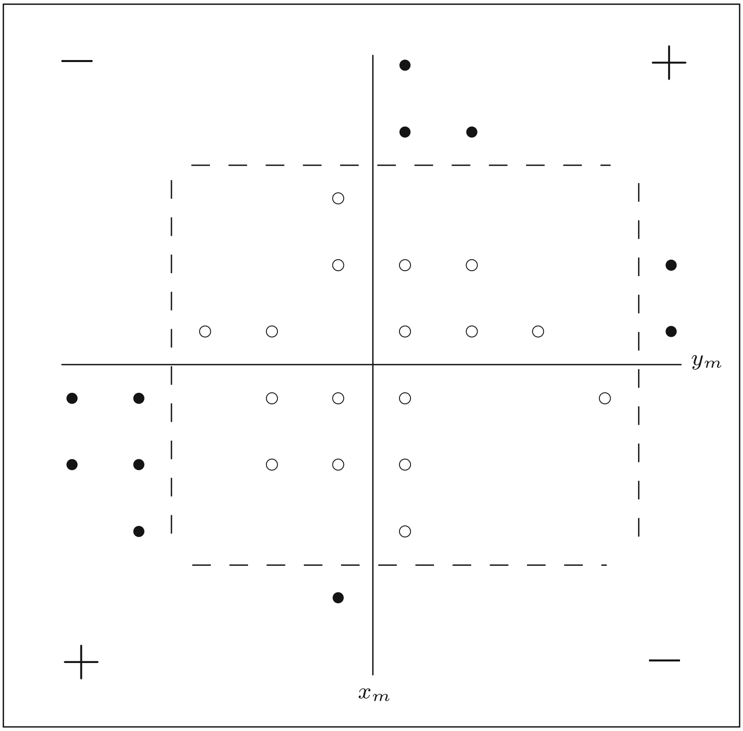


Fig. 3.10 Scatter diagram of $n = 28$ pairs of observations to illustrate the corner test of association by P.S. Olmstead and J.W. Tukey [1059]

3.15 Haldane–Smith and a Test for Birth-Order Effects

In 1948 John Haldane and Cedric Smith proposed a recursively-obtained two-sample rank-sum test for birth-order effects that employed a clever decomposition procedure similar to that used by Festinger in 1946 [573].

J.B.S. Haldane

John Burton Sanderson Haldane was educated at Eton and New College, University of Oxford, and was a commissioned officer during World War I. At the conclusion of the war, Haldane was awarded a fellowship at New College, University of Oxford, and then accepted a readership in biochemistry at Trinity College, University of Cambridge. In 1932 Haldane was elected Fellow of the Royal Society and a year later, became Professor of Genetics at University College, London. In the 1930s Haldane joined the Communist

(continued)

Party and assumed editorship of the party's London Paper, the *Daily Worker*. In 1956 Haldane, disillusioned with the official Party line and the rise of the Soviet biologist Trofim Lysenko, immigrated to India where he joined the Indian Statistical Institute at the invitation of P.C. Mahalanobis. In 1961 he resigned from the Indian Statistical Institute and accepted a position as Director of the Genetics and Biometry Laboratory in Orissa, India. Haldane wrote 24 books, including science fiction and stories for children, more than 400 scientific research papers, and innumerable popular articles [869]. John Burton Sanderson Haldane F.R.S. died of cancer on 1 December 1964, whereupon he donated his body to Rangaraya Medical College, Kakinada, India [869].

C.A.B. Smith

Cedric Austen Bardell Smith attended University College, London. In 1935, Smith received a scholarship to Trinity College, University of Cambridge, where he earned his Ph.D. in 1942.[39] In 1946 Smith was appointed Assistant Lecturer at the Galton Biometric Laboratory, University College, London, where he first met Haldane. In 1964 Smith accepted an appointment as the Weldon Professor of Biometry at University College, London. Cedric Smith clearly had a sense of humor and was known to occasionally sign his correspondence as "U.R. Blanche Descartes, Limit'd," which was an anagram of Cedric Austen Bardell Smith [1008]. Smith contributed to many of the classical topics in statistical genetics, including segregation ratios in family data, kinship, population structure, assortative mating, genetic correlation, and estimation of gene frequencies [1008]. Cedric Austen Bardell Smith died on 10 January 2002, just a few weeks shy of his 85th birthday [400, 1008].

In 1948 Haldane and Smith introduced an exact test for birth-order effects [573]. They had previously observed that in a number of hereditary diseases and abnormalities, the probability that any particular member of a sibship had a specified abnormality depended in part on his or her birth rank (birth order) [573, p. 117]. The test they proposed was based on the sum of birth ranks of all affected cases in all sibships. In a classic description of an exact permutation test, Haldane and Smith noted that if in each sibship the numbers of normal and affected siblings were

[39] Cedric Smith, Roland Brooks, Arthur Stone, and William Tutte met at Trinity College, University of Cambridge, and were known as the Trinity Four. Together they published mathematical papers under the pseudonym Blanche Descartes, much in the tradition of the putative Peter Ørno, John Rainwater, and Nicolas Bourbaki.

held constant, then if birth rank had no effect, every possible order of normal and affected siblings would be equally-probable. Accordingly, the sum of birth ranks for affected siblings would have a definite distribution, free from unknown parameters, providing "a 'conditional' and 'exact' test for effect of birth-rank" [573, p. 117]. Finally, they observed that this distribution would be very nearly normal in any practically occurring case with a mean and variance that were easily calculable.

Consider a single sibship of k births, h of which are affected. Let the birth ranks of the affected siblings be denoted by $a_1, a_2, \ldots, a_h$ and their sum by $A = \sum_{r=1}^{h} a_r$. Then, there are

$$\binom{k}{h} = \frac{k!}{h!(k-h)!} \tag{3.5}$$

equally-likely ways of distributing the h affected siblings.[40] Of these, the number of ways of distributing them, $P_{h,k}(A)$, so that their birth ranks sum to A is equal to the number of partitions of A into h unequal parts, $a_1, a_2, \ldots, a_h$, no part being greater than k. Given this, the probability $p_{h,k}(A)$ of obtaining a sum A is given by

$$p_{h,k}(A) = P_{h,k}(A) \Bigg/ \binom{k}{h} \ . \tag{3.6}$$

Dividing these partitions into two classes according to whether the greatest part is or is not k, yields

$$P_{h,k}(A) = P_{h,k-1}(A) + P_{h-1,k-1}(A-k) \ . \tag{3.7}$$

Haldane and Smith observed that from the relation described in Eq. (3.7) they could readily calculate $P_{h,k}(A)$ for small samples of h and k.

Since $(k+1-a_1), (k+1-a_2), \ldots, (k+1-a_h)$ must be a set of h integers, all different and not greater than k, and summing to $h(k+1)-A$, they showed that

$$P_{h,k}(A) = P_{h,k}[h(k+1)-A] \ . \tag{3.8}$$

Haldane and Smith went on to note that, similar to the affected siblings, in any sibship the unaffected siblings would all have different birth ranks, none exceeding k, but summing to $k(k+1)/2 - A$. Thus,

$$P_{h,k}(A) = P_{k-h,k}\,[k(k+1)/2 - A] \ . \tag{3.9}$$

An example will serve to illustrate the recursion procedure employed by Haldane and Smith.[41] Consider a sibship of $k = 6$ siblings with $h = 2$ of the siblings

[40] Equation (3.5) is incorrect in Haldane and Smith [573, p. 117] and is corrected here.

[41] It should be noted that while the decomposition in Eq. (3.8) is different from that employed by Mann and Whitney in Eq. (3.4) [880], it is similar to the decomposition used by Festinger [427], although there is no indication that Haldane and Smith were familiar with the work of Festinger.

Table 3.8 Partitions (P), sums (A), and frequencies (f) for $P_{h,k}(A) = P_{2,6}(7)$, $P_{h,k-1}(A) = P_{2,5}(7)$, $P_{h-1,k-1}(A-k) = P_{1,5}(1)$, and $P_{k-h,k}[k(k+1)/2 - A] = P_{4,6}(14)$

$P_{2,6}(7)$			$P_{2,5}(7)$			$P_{1,5}(1)$			$P_{4,6}(14)$		
P	A	f	P	A	f	P	A	f	P	A	f
1, 2	3	1	1, 2	3	1	1	1	1	1, 2, 3, 4	10	1
1, 3	4	1	1, 3	4	1	2	2	1	1, 2, 3, 5	11	1
1, 4	5	2	1, 4	5	2	3	3	1	1, 2, 3, 6	12	2
1, 5	6	2	1, 5	6	2	4	4	1	1, 2, 4, 5	13	2
1, 6	7	3	2, 3	7	2	5	5	1	1, 2, 4, 6	14	3
2, 3	8	2	2, 4	8	1				1, 2, 5, 6	15	2
2, 4	9	2	2, 5	9	1				1, 3, 4, 5	16	2
2, 5	10	1	3, 4						1, 3, 4, 6	17	1
2, 6	11	1	3, 5						1, 3, 5, 6	18	1
3, 4			4, 5						1, 4, 5, 6		
3, 5									2, 3, 4, 5		
3, 6									2, 3, 4, 6		
4, 5									2, 3, 5, 6		
4, 6									2, 4, 5, 6		
5, 6									3, 4, 5, 6		

classified as affected (a) and $k - h = 6 - 2 = 4$ of the siblings classified as normal (n), with birth order indicated by subscripts: n_1, a_2, n_3, n_4, a_5, n_6. Thus, the affected siblings are the second and fifth born out of six siblings and yield a sum of $A = a_2 + a_5 = 2 + 5 = 7$. Table 3.8 lists the partitions and associated frequency distributions for $h = 2$ and $k = 6$ in the first set of columns, $h = 2$ and $k - 1 = 6 - 1 = 5$ in the second set of columns, $h - 1 = 2 - 1 = 1$ and $k - 1 = 6 - 1 = 5$ in the third set of columns, and $k - h = 6 - 2 = 4$ and $k = 6$ in the fourth set of columns. It can be seen in Table 3.8 that $P_{h,k}(A) = P_{2,6}(7) = 3$ since there are three ways of placing an affected sibling yielding a sum of $A = 7$, i.e., $\{1,\ 6\}$, $\{2,\ 5\}$, and $\{3,\ 4\}$. As there are a total of

$$\binom{k}{h} = \frac{k!}{h!(k-h)!} = \frac{6!}{2!(6-2)!} = 15$$

equally-probable ways of placing the $h = 2$ affected siblings, the probability of obtaining a sum of $A = 7$ as given in Eq. (3.6) is

$$p_{2,6}(7) = P_{2,6}(7)/15 = 3/15 = 0.20 \ .$$

Dividing the partitions into two classes as in Eq. (3.7) yields

$$\begin{aligned} P_{2,6}(7) &= P_{2,6-1}(7) + P_{2-1,6-1}(7-6) \ , \\ 3 &= P_{2,5}(7) + P_{1,5}(1) \ , \\ 3 &= 2 + 1 \ , \end{aligned}$$

as illustrated in Table 3.8, where $P_{2,6}(7)$ in the first set of columns is associated with a frequency of 3, $P_{2,5}(7)$ in the second set of columns is associated with a frequency of 2, and $P_{1,5}(7-6) = P_{1,5}(1)$ in the third set of columns is associated with a frequency of 1; thus, $3 + 2 + 1$. Note that once again, the decomposition observed in the discussion of Festinger and the two-sample rank-sum test appears wherein

$$\begin{aligned}\binom{k}{h} &= \binom{k-1}{h} + \binom{k-1}{h-1} , \\ \binom{6}{2} &= \binom{6-1}{2} + \binom{6-1}{2-1} , \\ \binom{6}{2} &= \binom{5}{2} + \binom{5}{1} , \\ 15 &= 10 + 5 .\end{aligned}$$

This decomposition can be observed in Table 3.8 where the column of frequencies for $P_{2,6}(A)$ in the first set of columns sums to 15, the column of frequencies for $P_{2,5}(A)$ in the second set of columns sums to 10, the column of frequencies for $P_{1,5}(A-k)$ in the third set of columns sums to 5, and $15 = 10 + 5$.

The affected siblings, $(k + 1 - a_2)$ and $(k + 1 - a_5)$, constitute a set of $h = 2$ integer values where $(6+1-2) = 5$ and $(6+1-5) = 2$ are all different with none greater than $k = 6$. The values 5 and 2 sum to $h(k + 1) - A = 2(6 + 1) - 7 = 7$. Thus, as in Eq. (3.8),

$$P_{2,6}(7) = P_{2,6}[2(6 + 1) - 7] = P_{2,6}[7] = 3 .$$

The first set of columns in Table 3.8 lists the partitions and frequency distribution of the partitions of $P_{2,6}(A)$ in which the sum $A = 7$ has a frequency of 3 based on the partitions of $\{1, 6\}$, $\{2, 5\}$, and $\{3, 4\}$.

On the other hand, the normal siblings, $(k + 1 - n_1)$, $(k + 1 - n_3)$, $(k + 1 - n_4)$, and $(k + 1 - n_6)$, constitute a set of $k - h = 6 - 2 = 4$ integer values where $(6 + 1 - 1) = 6$, $(6 + 1 - 3) = 4$, $(6 + 1 - 4) = 3$, and $(6 + 1 - 6) = 1$ are all different with none greater than $k = 6$. The values 6, 4, 3, and 1 sum to $k(k + 1)/2 - A = 6(6 + 1)/2 - 7 = 14$. Thus, as in Eq. (3.9),

$$P_{2,6}(7) = P_{6-2,6}\,[6(6 + 1)/2 - 7] = P_{4,6}[14] = 3 .$$

The rightmost set of columns in Table 3.8 lists the partitions and frequency distribution of $P_{4,6}[k(k + 1)/2 - A]$ in which the sum $A = 14$ has a frequency of 3 based on the partitions $\{1, 2, 5, 6\}$, $\{1, 3, 4, 6\}$, and $\{2, 3, 4, 5\}$.

From Eqs. (3.8) and (3.9), Haldane and Smith were able to construct a table of values of $P_{h,k}(A)$ and $\binom{k}{h}$, giving the exact distribution for all values of k up to and including 12, noting that values not explicitly given in the table could readily be derived by the use of Eqs. (3.8) and (3.9). Additionally, Haldane and Smith investigated the approximate distribution of A. They found it more efficient to test $6A$ instead of A and showed that the theoretical mean of $6A$ was $3h(k+1)$ and the theoretical variance was $3h(k+1)(k-h)$, and thus provided a table of means and variances for $h = 1, \ldots, 18$ and $k = 2, \ldots, 20$. They observed that since A is made up of a number of independent components, the distribution of A would be approximately normal and, therefore, if an observed value of A exceeded the mean by more than twice the standard deviation, siblings born later were most likely to be affected, but if the observed value of A fell short of the mean by the same amount, siblings born earlier were most likely to be affected [573, p. 121]. They concluded the paper with an example analysis based on data from T.A. Munro on phenylketonuria from forty-seven British families that had previously been published in *Annals of Human Genetics* in January of 1947 [1014].[42]

3.16 Finney and the Fisher–Yates Test for 2 × 2 Tables

In 1948 David Finney constructed and published tables of exact probability values based on the hypergeometric distribution for testing the significance of data arranged in a 2 × 2 contingency table [434].

D.J. Finney

David John Finney read mathematics and statistics at Clare College, University of Cambridge, from 1934 to 1938. During his second year at Cambridge, doctors found a small spot on one lung and suggested he move to England's south coast for a brief period during the summer to recuperate. While there, he developed typhoid fever and was hospitalized for weeks, which caused him to miss an entire term at Cambridge. Finney later recalled that the fever had been a "happy accident" because it allowed him to deliberate on his future and rethink his plans to become a mathematician. Thus, when Finney returned to the University of Cambridge in 1937, he took the advice of an advisor to try his hand at statistics and signed up to take a course from John Wishart. It was Wishart who later told Finney about a competitive post-graduate opportunity

(continued)

[42] Phenylketonuria (PKU) is a autosomal recessive metabolic genetic disorder that can lead to mental retardation, seizures, behavioral problems, and autism. Dr. Asbjørn Følling, a Norwegian biochemist and physician, was the first to publish a description of phenylketonuria as a cause of mental retardation in 1934 [475].

Fig. 3.11 Notation for the Finney standard 2 × 2 contingency table

a	$A-a$	A
b	$B-b$	B
$a+b$	$A+B-a-b$	$A+B$

that took him to the Galton Biometric Laboratory at University College, London, to work with R.A. Fisher.

In 1939 Finney accepted a position as assistant to Frank Yates at Rothamsted Experimental Station to replace William G. Cochran who was leaving to assume a post at Iowa State University. After leaving Rothamsted, Finney taught statistics at the University of Oxford, the University of Aberdeen, and the University of Edinburgh. Finney was elected Fellow of the Royal Society in 1955 and was president of the Royal Society in 1973. Finney retired from the University of Edinburgh in 1984 but has continued researching, with a focus on drug safety [435, 866] David John Finney F.R.S. was born on 3 January 1917 and at the time of this writing is 96 years old.

In 1948 Finney considered the Fisher–Yates exact permutation test[43] of significance for 2 × 2 contingency tables [434]. Acknowledging that the usual chi-squared test of significance was questionable when the expected cell frequencies were small, Finney utilized exact hypergeometric probability values to construct a table of significance levels for 2 × 2 contingency tables with small expected frequencies. Thus, as Finney explained, for a standard 2 × 2 contingency table with cell frequencies and marginal frequency totals represented as in Fig. 3.11, the hypergeometric probability for cell b, with fixed marginal frequency totals, is given by

$$P\{b|B,\, a+b,\, A+B\} = \frac{A!\,B!\,(a+b)!\,(A+B-a-b)!}{(A+B)!} \times \frac{1}{a!\,b!\,(A-a)!\,(B-b)!}\;. \tag{3.10}$$

Note that the first factor to the right of the equal sign in Eq. (3.10) is dependent only on the five marginal frequency totals, while the second factor to the right of the equal sign depends only on the four internal cell frequencies. The table presented by Finney enabled tests of significance at one-tailed probability levels of $\alpha = 0.05$, 0.025, 0.01, and 0.005, to be made by direct reference for any 2 × 2 contingency

[43] The Fisher–Yates test of significance for 2 × 2 contingency tables was independently developed by R.A. Fisher in 1935 [452], F. Yates in 1934 [1472], and J.O. Irwin in 1935 [674] (qq.v. pages 25, 37, and 48).

table having no marginal frequency total greater than 15.[44] Finney illustrated the use of the table of significance levels with Johannes Lange's data on criminal behavior among twin brothers or sisters of criminals, previously analyzed by R.A. Fisher (q.v. page 41).

The table to test significance published by Finney in 1948 was limited to 2 × 2 contingency tables with both the marginal frequency totals in either the rows or columns less than or equal to 15. Latscha [804] extended Finney's table with tables containing marginal frequency totals up to 20 in 1953, and Armsen [34] further extended Latscha's tables to marginal frequencies up to 50 in 1955.

3.17 Lehmann–Stein and Non-parametric Tests

In 1949, in a highly theoretical article published in *The Annals of Mathematical Statistics*, Erich Lehmann and Charles Stein investigated optimum tests for non-parametric hypotheses against certain classes of alternatives [818].

E.L. Lehmann

Erich Leo Lehmann studied mathematics at Trinity College, University of Cambridge, before moving to the University of California at Berkeley as a graduate student in 1940, where he was surprisingly admitted without having earned an undergraduate degree. Lehmann received his M.A. degree from Berkeley in 1942, followed by his Ph.D. in 1946, both in mathematics and both under Jerzy Neyman (q.v. page 21). After brief teaching assignments at Columbia University, Princeton University, and Stanford University, Lehmann returned to Berkeley in 1952 as an Associate Professor. In 1954 he was appointed Professor of Mathematics and the following year, Berkeley formed a Statistics Department at which time Lehmann became a Professor of Statistics. Lehmann remained at Berkeley for the remainder of his academic career, retiring in 1988. Retirement did not mean, however, that Lehmann ceased working. In fact, Lehmann completed work on his last book, *Fisher, Neyman, and the Creation of Classical Statistics*, in 2009. The book was published posthumously by Springer in 2011. Erich Leo Lehmann died at home in Berkeley on 12 September 2009 at the age of 91 [38, 215, 337, 1187].

[44] Unfortunately, Finney recommended doubling the obtained one-tailed probability value when using a two-tailed test [434, p. 146]. This was destined to become a procedure of considerable controversy in the mid-1980s (q.v. page 51).

C.M. Stein

Charles M. Stein earned his B.S. in mathematics from the University of Chicago in 1940 and began graduate work at Chicago, but his graduate studies were interrupted by military service during World War II. After leaving the Air Force in 1946, Stein moved to Columbia University, earning his Ph.D. in mathematical statistics under Abraham Wald (q.v. page 122) in 1947. Upon graduation, Stein worked first at the Neyman Statistics Laboratory at the University of California at Berkeley and then from 1951 to 1953 was an Associate Professor at the University of Chicago. Stein joined the faculty at Stanford University in 1953, where he remained for the rest of his academic career. Stein retired in 1989 and in 2010, Stanford held a symposium in probability and statistics in honor of Stein's 90th birthday [336,751,767,814]. Charles Stein was born on 22 March 1920 and at the time of this writing is 93 years old.

In a 1949 article Eric Lehmann and Charles Stein researched permutation tests in a very general framework. Let $Z_1, \ldots, Z_N = Z$ denote N random variables and suppose there is a partition of the sample space $z_1, \ldots, z_N = z$ into classes of equivalent points. Denote by T_z the set of all points that are equivalent to z, which contains a finite number of points, r, and let H be the hypothesis that the distribution of Z is, for any z, invariant over all the points in T_z. Then a test of H is a function of φ that assigns to each point z a number φ_z between zero and one representing the probability of rejecting H when z is observed. If

$$\sum_{z' \in T_z} \varphi(z') = \alpha r$$

identically in z, then φ is a similar size-α test of statistic H. Lehmann and Stein showed that a most powerful and similar size-α test of H against a simple alternative is given by ordering the points of T_z so that

$$u\left(z^{(1)}\right) \geq \cdots \geq u\left(z^{(r)}\right)$$

and setting

$$\varphi(z) = \begin{cases} 1 & \text{if } u\,(z) > u\left(z^{(1+[\alpha r])}\right) , \\ \alpha & \text{if } u\,(z) = u\left(z^{(1+[\alpha r])}\right) , \\ 0 & \text{if } u\,(z) < u\left(z^{(1+[\alpha r])}\right) , \end{cases}$$

where u is an appropriately chosen function and $\alpha = \alpha(z)$ is uniquely determined to provide a size-α test [818].

Lehmann and Stein stated that in many experimental situations, the hypothesis that the distribution of the Zs was invariant under all permutations was more realistic than the hypothesis that the Zs were independently and identically distributed. They also noted in a discussion of alternative hypotheses that many of the alternative hypotheses considered, for example those involving normality, were dictated more by tradition and ease of treatment than by appropriateness in actual experiments [818, p. 29].

3.18 Rank-Order Statistics

The years 1948–1950 constituted a defining period for rank-order statistical methods. The year 1948 saw the publication of M.G. Kendall's deceptively small 160 page volume on *Rank Correlation Methods* [734]; also in 1948, a massive summary of order statistics by S.S. Wilks was published in *Bulletin of the American Mathematical Society* [1456]. In March 1950, a special symposium on ranking methods was held by the Research Section of the Royal Statistical Society and chaired by M.G. Kendall, with presenters that included P.A.P. Moran, J.W. Whitfield, and H.E. Daniels, along with several discussants, including R.L. Plackett, B. Babington Smith, A. Stuart, J.I. Mason, I.J. Good, S.T. David, and L.T. Wilkins. The text of the symposium was later published in *Journal of the Royal Statistical Society, Series B* [314, 1005, 1444]. Although the presentations by Moran, Whitfield, and Daniels contained little on permutation methods per se, Kendall's book was replete with discussions of permutation statistics and the article by Wilks constituted a rich source on permutation methods for its time [1456].

3.18.1 Kendall and Rank Correlation Methods

The importance of Kendall's 1948 book on rank-order correlation methods cannot be overstated, as it forever changed the field of rank-order statistics (q.v. page 84). It has gone through five editions, the last edition with J.D. Gibbons, it has been cited over 5,000 times, and it is still in print. The title of Kendall's book, *Rank Correlation Methods*, is perhaps a little misleading as it contained much more than rank-order correlation methods, including an extensive summary of permutation methods. Of particular relevance to permutation methods, Kendall included descriptive summaries of articles that contained permutation statistics per se and tables of exact probability values obtained from permutation distributions.

For example, Kendall summarized articles by H. Hotelling and M.R. Pabst that used permutation methods for calculating exact probability values for small samples of ranked data in their research on simple bivariate correlation [653]; E.J.G. Pitman on permutation tests for two independent samples, bivariate correlation, and randomized blocks analysis of variance [1129–1131]; M. Friedman on procedures employing ranked data in place of the ordinary analysis of variance

[485]; M.G. Kendall on exact probability values for the τ_b measure of rank-order correlation [728]; E.G. Olds on exact probability values for Spearman's rank-order correlation coefficient [1054]; B.L. Welch on exact probability values for the η^2 test of homogeneity [1429]; M.G. Kendall and B. Babington Smith on exact probability values for the coefficient of consistency [741]; H.B. Mann on tables of exact probability values for tests of randomness against trend [879]; F. Wilcoxon on tables of exact probability values for the two-sample test for rank-order statistics [1453]; and H.B. Mann and D.R. Whitney on exact probability values for the two-sample rank-sum test [880].[45]

3.18.2 Wilks and Order Statistics

In 1948 Samuel S. Wilks of Princeton University published a lengthy article on order statistics in *Bulletin of the American Mathematical Society* that summarized contributions by a large number of statisticians on a comprehensive collection of statistical tests and measures and included an exhaustive list of references [1456].

S.S. Wilks

Samuel Stanley Wilks earned his B.A. degree in industrial arts at North Texas State Teachers College (now, the University of North Texas) in 1926, his M.A. degree in mathematics at the University of Texas, and his Ph.D. in statistics at the University of Iowa in 1931. Upon graduation with his Ph.D., Wilks was awarded a National Research Council Fellowship in mathematics at Columbia University, where he studied with Harold Hotelling. In 1932 Wilks was appointed as a National Research Council International Fellow and studied at both the University of London and the University of Cambridge. There, Wilks had the opportunity to work with both Karl Pearson and John Wishart.

In 1934, at the recommendation of Harold Hotelling, Wilks was recruited to Princeton University by the Chair of the Department of Mathematics, Luther Pfahler Eisenhart, who was the father of Churchill Eisenhart by his first wife. Later, Churchill Eisenhart would earn his M.A. degree under Wilks. Wilks remained at Princeton for his entire career. Samuel Stanley Wilks died unexpectedly in 1964 as lamented in the opening sentences of his obituary by Frederick Mosteller:

(continued)

[45] It should be noted that Kendall neglected to mention the two-sample rank-sum test developed by Festinger 2 years prior, perhaps because it was published in the psychology journal, *Psychometrika*, which was not commonly read by statisticians.

> [t]he death in his sleep of Samuel Stanley Wilks at his Princeton home on March 7, 1964, ended a life of dedicated service to statistics, education, and the nation. Apparently in the best of health, his sudden death at the age of 57 shocked and saddened the entire statistical community [1012, p. 411].

While at Princeton, Wilks was the editor of *Annals of Mathematical Statistics* from 1938 to 1949. In addition, for 30 years Wilks worked with the College Entrance Examination Board (CEEB) and with the Educational Testing Service (ETS), advising on research design and analysis, score scaling, the development of mathematical tests, and studies of mathematical education [325, 692, 814, 1012].

Kendall's *Rank Correlation Methods* was quickly followed by a substantial and sophisticated exposition of order statistics by S.S. Wilks in 1948 [1456]. In a highly structured organization, Wilks provided a lengthy discourse on order statistics, summarizing the results on order statistics, and listing all the references up to that time. Although the title of the article was "Order statistics," the article was also a rich source on permutation methods.

This article by Wilks on order statistics comprised some 45 pages in *Bulletin of the American Mathematical Society* and is too extensive to be summarized completely here. The article included summaries of the contributions to permutation methods by R.A. Fisher on permutation tests in general [448, 451]; H. Hotelling and M.R. Pabst on exact probability values for ranked data [653]; M. Friedman on the analysis of variance for ranks [485]; E.J.G. Pitman's classic three articles on permutation versions of the two-sample test, bivariate correlation, and randomized blocks analysis of variance [1129–1131]; B.L. Welch on permutation tests for randomized block and Latin square designs [1428]; E.G. Olds on a permutation approach to rank-order correlation [1054]; W.J. Dixon on a permutation approach to a two-sample test [353]; A.M. Mood on the exact distribution of runs [999]; H. Scheffé's seminal article on non-parametric statistical inference [1230]; F.S. Swed and C. Eisenhart on exact probability values for the runs test [1337]; A. Wald and J. Wolfowitz on two-sample tests and serial correlation [1405, 1406]; and P.S. Olmstead and J.W. Tukey on exact probability values for the quadrant-sum test [1059].

3.19 van der Reyden and a Two-Sample Rank-Sum Test

In 1952 D. van der Reyden proposed a two-sample rank-sum test that was equivalent to those published previously by Wilcoxon in 1945, Festinger in 1946, Mann and Whitney in 1947, Whitfield in 1947, and Haldane and Smith in 1948. However, the approach was quite different, as it was based on a novel tabular procedure [1391].

D. van der Reyden

Little is known about Dirk van der Reyden other than that early in the 1950s he was a statistician for the Tobacco Research Board in Salisbury, Southern Rhodesia.[46] In 1957 he earned a Ph.D. in experimental statistics from North Carolina State University at Raleigh and then joined the faculty at Washington University in St. Louis. In 1952 van der Reyden independently developed a two-sample rank-sum test equivalent to the tests of Wilcoxon [1453], Festinger [427], Mann and Whitney [880], Whitfield [1443], and Haldane and Smith [573], although none of these is referenced; in fact, the article by van der Reyden contains no references whatever. The stated purpose of the proposed test was to provide a simple exact test of significance using sums of ranks in order to avoid computing sums of squares [1391, p. 96].

In a novel approach, van der Reyden utilized a tabular format involving rotations of triangular matrices to generate permutation distribution frequencies and published tables of critical values at two-tailed significance levels of 0.05, 0.02, and 0.01 for all sample sizes such that if m and n denote the population and sample sizes, respectively, $10 \leq m \leq 30$ and $2 \leq n \leq 12$ at the 0.05 level, and $3 \leq n \leq 12$ at the 0.02 and 0.01 levels [1391]. This work went largely unnoticed for some years, appearing as it did in the relatively obscure *Rhodesia Agricultural Journal*.

In 1952 D. van der Reyden proposed a tabular procedure for the two-sample rank-sum test that was equivalent to tests previously proposed by Wilcoxon in 1945 [1453], Festinger in 1946 [427], Mann and Whitney in 1947 [880], Whitfield in 1947 [1443], and Haldane and Smith in 1948 [573].[47] Table 3.9 illustrates the van der Reyden tabular procedure with values of $n = 1, 2, 3$, $m = 1, \ldots, 6$, and sums of frequencies from $T = 1$ to $T = 15$. Looking first at the column headed $n = 1$ in Table 3.9, note that when $m = 1$ and $n = 1$, $T = 1$; when $m = 2$ and $n = 1$, $T = 1$ or 2; when $m = 3$ and $n = 1$, $T = 1$, 2, or 3; and when $m = 4$ and $n = 1$, $T = 1$, 2, 3, or 4. Simply put, taking all samples of one item from m items, all values of T will have a frequency of 1. In this case, each T has a frequency of 1 and each frequency sums to $\binom{m}{n}$, e.g., for $m = 4$ and $n = 1$ the frequency distribution is $\{1, 1, 1, 1\}$ with a sum of 4, which is $\binom{4}{1} = 4$. To obtain the frequencies for samples of $n = 2$ items, rotate all frequencies for $n = 1$ clockwise through 45°, shifting the whole distribution downward to

[46] Southern Rhodesia was shortened to Rhodesia in 1965 and renamed the Republic of Zimbabwe in 1980.

[47] For a brief history of the development of the two-sample rank-sum test, see a 2012 article by Berry, Mielke, and Johnston in *Computational Statistics* [160].

Table 3.9 Generation of frequency arrays for $n = 1$, $n = 2$, and $n = 3$ as described by van der Reyden [1391]

	$n = 1$				$n = 2$								$n = 3$							
T/m	1	2	3	4	2	3	4	5	2	3	4	5	3	4	5	6	3	4	5	6
1	1	1	1	1																
2		1	1	1																
3			1	1	1				1	1	1	1								
4				1		1				1	1	1								
5						1	1			1	2	2								
6							1	1			1	2	1				1	1	1	1
7							1	1			1	2		1				1	1	1
8								1				1		1	1			1	2	2
9								1				1		1	1	1		1	2	3
10															2	1			2	3
11															1	2			1	3
12															1	2			1	3
13																2				2
14																1				1
15																1				1

$$T = \binom{n+1}{2} = \frac{n(n+1)}{2} \, .$$

Thus in Table 3.9, the frequencies obtained for $n = 1$ are transposed with the first row now constituting the fourth column, the second row constituting the third column, and so on. Then this transposed matrix is shifted downward so that it begins at $T = n(n+1)/2 = 2(2+1)/2 = 3$. Finally, the frequencies are added together horizontally in a specific manner, as follows.

Consider the frequency distributions listed under $n = 2$ in Table 3.9. There are two sets of frequency distributions under $n = 2$, one on the left and one on the right, both labeled $m = 2, 3, 4, 5$. So, for example, to create the frequency distribution listed under $n = 2$, $m = 3$ on the right, add together the frequency distribution listed under $n = 2$, $m = 2$ on the right and the frequency distribution under $n = 2$, $m = 3$ on the left. To create the frequency distribution listed under $n = 2$, $m = 4$ on the right, add together the frequency distribution listed under $n = 2$, $m = 3$ on the right and the frequency distribution under $n = 2$, $m = 4$ on the left. To create the frequency distribution listed under $n = 2$, $m = 5$ on the right, add together the frequency distribution listed under $n = 2$, $m = 4$ on the right and the frequency distribution under $n = 2$, $m = 5$ on the left. The process continues in this manner, recursively generating the required frequency distributions.[48]

[48] Authors' note: in deciphering the article by van der Reyden we were often reminded of a comment by Nathaniel Bowditch. In the memoir prefixed to the fourth volume of Bowditch's translation of Laplace's *Mécanique Céleste*, page 62, Bowditch wrote: "[w]henever I meet in La

For a final example, consider the frequency distributions listed under $n = 3$. Again there are two sets of frequency distributions, one on the left and one on the right. The distribution on the left is created by rotating the distribution created under $n = 2$ on the right, and shifting it downward so it begins at $T = n(n+1)/2 = 3(3+1)/2 = 6$. To create the frequency distribution listed under $n = 3$, $m = 6$ on the right, add together the frequency distribution listed under $n = 3$, $m = 5$ on the right and the frequency distribution under $n = 3$, $m = 6$ on the left. The frequency distributions of sums in Table 3.9 can be compared with the frequency distributions of sums in Tables 3.4 and 3.5 that were generated with Festinger's method [427]. In this recursive manner, van der Reyden created tables for T from $n = 2, \ldots, 12$ and $m = 10, \ldots, 30$ for the $\alpha = 0.05$, 0.02, and 0.01 levels of significance.

3.20 White and Tables for the Rank-Sum Test

Although trained as a medical doctor, Colin White also contributed to the field of permutation statistics. In 1952 White recursively generated tables of exact probability values for the Wilcoxon two-sample rank-sum test in which the sample sizes could either be equal or unequal [1441].

C. White

Colin White earned his M.S. and his M.D. degrees from the University of Sydney, Australia, in 1937 and 1940, respectively. Upon graduation, White served as a medical officer for the Commonwealth Department of Health in Canberra, then moved to England where he was a lecturer at the University of Birmingham. White immigrated to the United States in 1948 and joined Yale University as an Assistant Professor in 1953. In 1962 he was promoted to Professor and, eventually, Chair of the Department of Epidemiology and Public Health. White retired in 1984, but continued his research as a senior research scientist at Yale University until 2007, enjoying a career that spanned six decades. Colin White passed away on 1 February 2011 at the advanced age of 97 [673].

In 1952 White introduced "elementary methods" to develop tables for the Wilcoxon two-sample rank-sum test when the numbers of items in the two

Place with the words 'Thus it plainly appears' I am sure that hours, and perhaps days of hard study will alone enable me to discover *how* it plainly appears." (Bowditch, quoted in Todhunter [1363, p. 478]; emphasis in the original).

independent samples, n_1 and n_2, were not necessarily equal [1441].[49] White provided three tables that gave critical values for rank sums for $n_1 = 2, \ldots, 15$ and $n_2 = 4, \ldots, 28$ for critical values of $\alpha = 0.05$, $n_1 = 2, \ldots, 15$ and $n_2 = 5, \ldots, 28$ for critical values of $\alpha = 0.01$, and $n_1 = 3, \ldots, 15$ and $n_2 = 7, \ldots, 27$ for critical values of $\alpha = 0.001$.

Following the notation of White, let n_1 denote the number of items in the sample for which the rank total, T, is required, and let n_2 represent the number of items in the second sample. The ranks to be allotted are $1, 2, \ldots, n_1 + n_2$, where the lowest value the rank total can have is given by

$$\frac{n_1(n_1+1)}{2} \, ,$$

the largest total is given by

$$\frac{n_1(n_1+2n_2+1)}{2} \, ,$$

and all integer values between these two limits are possible rank totals. For example, consider $n_1 = 5$ items drawn from the consecutive integers $1, 2, \ldots, 12$, where the lowest rank total is

$$\frac{n_1(n_1+1)}{2} = \frac{5(5+1)}{2} = 1+2+3+4+5 = 15 \, ,$$

and the highest rank total is

$$\frac{n_1(n_1+2n_2+1)}{2} = \frac{5[5+(2)(7)+1]}{2} = 8+9+10+11+12 = 50 \, .$$

White's recursion procedure to obtain rank-sum totals described here is similar to Wilcoxon's procedure [1453]. Let $W_T^{n_1,n_2}$ denote the number of ways of obtaining a rank total when there are n_1 items in the sample of which T is required, and n_2 items in the second sample. Now, as White showed, $W_T^{n_1,n_2}$ can be obtained recursively; thus,

$$W_T^{n_1,n_2} = W_T^{n_1,(n_2-1)} + W_{T-n_1-n_2}^{(n_1-1),n_2} \, .$$

For example, as shown in Table 3.10 there are 18 ways of obtaining a total of $T = 23$ when $n_1 = 5$ of the integers $1, 2, 3, \ldots, n_1 + n_2 = 13$ are summed without repetitions, 17 ways of obtaining a total of $T = 23$ when $n_1 = 5$ of the integers $1, 2, 3, \ldots, n_1 + (n_2 - 1) = 12$ are summed, and only one way of obtaining a total of $T = 10$ when $n_1 = 4$ of the integers $1, 2, 3, \ldots, (n_1 - 1) + n_2 = 12$ are summed. Specifically, when $n_1 = 5$, $n_2 = 8$, and $T = 23$,

[49] Recall that the two-sample rank-sum method proposed by Wilcoxon in 1945 considered only equal sample sizes [1453] and Festinger, in 1946, was the first to develop a two-sample rank-sum procedure that could accommodate different sample sizes [427].

Table 3.10 Number of ways a sum of $T = 23$ can be obtained from 5 integers chosen from 13, a sum of $T = 23$ can be obtained from 5 integers chosen from 12, and a sum of $T = 10$ can be obtained from 4 integers chosen from 12

Count	5 from 13	5 from 12	4 from 12
1	2 3 5 6 7	2 3 5 6 7	1 2 3 4
2	1 4 5 6 7	1 4 5 6 7	
3	2 3 4 6 8	2 3 4 6 8	
4	1 3 5 6 8	1 3 5 6 8	
5	1 3 4 7 8	1 3 4 7 8	
6	1 2 5 7 8	1 2 5 7 8	
7	2 3 4 5 9	2 3 4 5 9	
8	1 3 4 6 9	1 3 4 6 9	
9	1 2 5 6 9	1 2 5 6 9	
10	1 2 4 7 9	1 2 4 7 9	
11	1 2 3 8 9	1 2 3 8 9	
12	1 3 4 5 10	1 3 4 5 10	
13	1 2 4 6 10	1 2 4 6 10	
14	1 2 3 7 10	1 2 3 7 10	
15	1 2 4 5 11	1 2 4 5 11	
16	1 2 3 6 11	1 2 3 6 11	
17	1 2 3 5 12	1 2 3 5 12	
18	1 2 3 4 13		

$$W_{23}^{5,\,8} = W_{23}^{5,\,(8-1)} + W_{23-5-8}^{(5-1),\,8}$$

and 18 = 17 + 1. Table 3.10 lists the various ways of obtaining a total of $T = 23$ from $W_{23}^{5,\,8}$, $W_{23}^{5,\,7}$, and a total of $T = 10$ from $W_{10}^{4,\,8}$. For sums based on 5 integers drawn from 13 consecutive integers, 18 out of a possible 1,287 sums equal $T = 23$; for sums based on 5 integers drawn from 12 consecutive integers, 17 out of a possible 792 sums equal $T = 23$; and for sums based on 4 integers drawn from 12 consecutive integers, only 1 out of a possible 495 sums equals $T = 10$. In this manner, White was able recursively to generate exact rank-sum totals for various combinations of n_1 and n_2.

3.21 Other Results for the Two-Sample Rank-Sum Test

In addition to the published tables already mentioned by Wilcoxon, Festinger, Mann and Whitney, White, and van der Reyden, several other extensions of tables of exact probability values appeared in the statistical literature during this period. Two are worth mentioning. In 1955, Fix and Hodges published extended tables for the Wilcoxon two-sample rank-sum W statistic [465].[50] If the sizes of the two samples

[50] For brief biographical sketches of Evelyn Fix and Joseph L. Hodges, see Lehmann's wonderful little book titled *Reminiscences of a Statistician: The Company I Kept*, published in 2008 [814, pp. 27–35].

are designated as m and n with $m \leq n$, the tables include exact probability values for $m \leq 12$. In 1953, Auble published extended tables for the Mann–Whitney two-sample rank-sum U statistic [40]. If n_1 and n_2 denote the sizes of the two samples, the tables from Auble give probability values for one- and two-sided tests for $\alpha = 0.05$ and $\alpha = 0.01$ for n_1 and n_2 from 1 to 20. In addition, many more tables of exact probability values appeared for W and U after 1959 when computers made generation of exact probability values much easier. Most notable among these were tables for the Wilcoxon two-sample rank-sum W statistic by Jacobson in 1963 [677] and extended tables for the Mann–Whitney two-sample rank-sum U statistic by Milton in 1964 [996].

Because the subject of interest is the historical development of permutation methods rather than a general discussion of statistics, much has had to be omitted in the discussion of the Wilcoxon and Mann–Whitney two-sample rank-sum tests. Consider that many of those who published tables of exact probability values also went on to provide approximate probability values for larger sample sizes. In general, they used methods based on moments to fit an approximate probability distribution. For example, in 1947 Wilcoxon provided tables of approximate probability values for both the unpaired and paired two-sample rank-sum tests [1454]. In addition, some of the published tables included adjustments for tied ranks, while some did not. Finally, there were errors in several of the published tables that were corrected in later articles; see especially the article by Verdooren [1398] that contained corrections for the tables by White [1441] and Auble [40], and an erratum to the article by Kruskal and Wallis [779] that contained corrections to the tables by White [1441] and van der Reyden [1391].

While these were all important contributions, they are not directly related to the focus on the structure and development of permutation statistical methods. However, a final note may be of some interest. The permutation methods to produce exact probability values introduced by Wilcoxon, Festinger, Mann and Whitney, Whitfield, Haldane and Smith, and van der Reyden are quite complex, but the test statistics W and U are relatively straightforward to compute. That does not mean, however, that they are simple to implement. In 2000, Bergmann, Ludbrook, and Spooren investigated the Wilcoxon–Mann–Whitney (WMW) procedures provided by eleven statistical packages. Some of the packages used large-sample approximations and some used exact permutation procedures. In the first case, some packages corrected for continuity and some did not. Moreover, some packages adjusted for tied ranks and some did not. Combinations of these choices led to very different results. The authors concluded that the "only infallible way of executing the WMW test is to compile the null distribution of the rank-sum statistic by exact permutation. This was…Wilcoxon's (1945) thesis and it provided the theoretical basis for his test" [100, p. 76].[51]

[51] In this regard, see an article by John Ludbrook on "The Wilcoxon–Mann–Whitney test condemned" in *British Journal of Surgery* [851] as well as a rejoinder by G.D. Murray [1017]. See also an exact permutation computer program for the Wilcoxon–Mann–Whitney test by Berry and Mielke in 2000 [155].

3.22 David–Kendall–Stuart and Rank-Order Correlation

In 1951 S.T. David, M.G. Kendall (q.v. page 84), and A. Stuart published an article concerning questions of distributions in the theory of rank-order correlation [328]. This article was motivated by three articles that had appeared the previous year in *Journal of the Royal Statistical Society*, the first by P.A.P. Moran on "Recent developments in ranking theory" [1005], the second by J.W. Whitfield on "Uses of the ranking method in psychology" [1444], and the third by H.E. Daniels on "Rank correlation and population models" [314]. Consequently, the article by David et al. is not primarily concerned with permutation methods. That said, this article does make a contribution of interest in a chronicle of permutation methods. David et al. noted that the exact distribution for Spearman's rank-order correlation coefficient had been given by Kendall, Kendall, and Babington Smith in 1939 for n, the rank number, from $n = 2$ to 8, inclusive [746], and independently by Olds (q.v. page 83) in 1938 for $n = 2$ to 7, inclusive [1054]. David et al. then proceeded to provide tables of the exact distribution of Spearman's rank-order correlation coefficient for $n = 9$ and $n = 10$. In 1955 Litchfield and Wilcoxon provided a table of critical totals of squared rank differences and a nomograph which permitted direct reading of the rank-order correlation coefficient for 6–40 pairs of observations and two probability levels, 0.05 and 0.01 [833].

What is of interest here are the comments by David et al. on the calculations of the exact distributions, as they reflect the difficulty in computing exact probability values in the years preceding the development of high-speed computers. David et al. observed that the method of obtaining the distributions used by both Kendall et al. in 1939 [746] and previously by Olds in 1938 [1054] were essentially the same, and further noted that "the work of explicit expansion rapidly increases as n becomes larger" [328, p. 131]. They went on to explain that they had been unable to find any methods of alleviating the amount of work required other than those methods previously described by Kendall et al. in 1939 and observed that the expansions to $n = 9$ and $n = 10$ were about as far as a computer's patience could be expected to extend.[52]

3.23 Freeman–Halton and an Exact Test of Contingency

In 1951 Gerald Freeman and John Halton published a short but influential article in *Biometrika* that addressed exact methods for analyzing two-way and three-way contingency tables, given fixed marginal frequency totals [480].

[52] It should be explained that "computer" was a common term that referred to the person who was responsible for calculations, usually a woman or group of women. In this case the computer was Miss Joan Ayling of the National Institute for Social and Economic Research who was given due credit by the authors "for her customary patience and accuracy" [328, p. 131]. See also a discussion by George Dyson in a 2012 book titled *Turing's Cathedral* [370, p. 59].

J.H. Halton

John H. Halton received his B.A. and M.A. degrees in mathematics and physics from the University of Cambridge in 1953 and 1957, respectively, and his Ph.D. from the University of Oxford in 1960. He held positions as a physicist in several locations, including the English Electric Company, the University of Oxford's Clarendon Laboratory, the University of Colorado in Boulder, the Brookhaven National Laboratory, and the University of Wisconsin at Madison. In 1984 he joined the Department of Computer Science at the University of North Carolina at Chapel Hill. In 2008 the University of Cambridge presented Halton with a D.Sc. degree, an honor that was based on forty of his published works.

In 1951 Gerald H. Freeman and John H. Halton published a short note on the exact treatment of contingency and goodness of fit. The purpose of the note was to present an exact method of analyzing r-way contingency tables with small cell frequencies to replace the chi-squared approximation that was considered unsuitable for small observed and expected values [480, pp. 141, 149]. The note is somewhat unique, as it contained no references to previous literature. The note, however, did include an exact treatment of $r \times c$ and $r \times c \times s$ contingency tables with fixed marginal frequency totals. The approach to the two-dimensional tables utilized the conventional hypergeometric probability distribution and can be illustrated with a 2×3 contingency table. Given fixed marginal frequency totals a_i, $i = 1, \ldots, r$, and b_j, $j = 1, \ldots, c$, let n denote the total number of objects and let x_{ij} denote a cell frequency for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. Finally, index each table by t. Then, the probability of the tth $r \times c$ contingency table is given by

$$P_{X^{(t)}} = \frac{\prod_{i=1}^{r} a_i! \prod_{j=1}^{c} b_j!}{n! \prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}^{(t)}!} ,$$

which had previously been put forward by R.A. Fisher in 1935 in the "lady tasting tea" experiment (q.v. page 58).[53] Freeman and Halton further defined

$$P_{X^{(t)}} = \frac{Q_L}{R_{X^{(t)}}} , \tag{3.11}$$

[53] A rigorous derivation of the exact contingency formula was given by John Halton in *Mathematical Proceedings of the Cambridge Philosophical Society* in 1969 [578].

where

$$Q_L = \frac{\prod_{i=1}^{r} a_i! \prod_{j=1}^{c} b_j!}{n!} \tag{3.12}$$

and

$$R_{X^{(t)}} = \prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}^{(t)}! \,. \tag{3.13}$$

As Freeman and Halton explained, by using logarithms the calculations could then be performed, with Q_L being determined once for all tables and $R_{X^{(t)}}$ separately for each table. For an example, consider an observed contingency table L given by

$$L = \begin{bmatrix} 0 & 3 & 2 \\ 6 & 5 & 1 \end{bmatrix} .$$

For any k-dimensional contingency table with fixed marginal frequency totals, $r_1 \times r_2 \times \cdots \times r_k$, there are v degrees of freedom, where

$$v = \prod_{m=1}^{k} r_m - \sum_{m=1}^{k} (r_m - 1) - 1 \,.$$

Since, for the k-dimensional contingency table, L, with $r_1 = 2$ and $r_2 = 3$, there are

$$v = (2)(3) - (2-1) + (3-1) - 1 = 2$$

degrees of freedom, only two independent cells need be manipulated, e.g., x_{11} and x_{12}, and the rest simply filled in, given the fixed marginal frequency totals. To illustrate, the first six of the possible 18 cell configurations are listed here, with the observed contingency table being $L = X^{(2)}$:

$$X^{(1)} = \begin{bmatrix} 0 & 2 & 3 \\ 6 & 6 & 0 \end{bmatrix} , \quad X^{(2)} = \begin{bmatrix} 0 & 3 & 2 \\ 6 & 5 & 1 \end{bmatrix} , \quad X^{(3)} = \begin{bmatrix} 0 & 4 & 1 \\ 6 & 4 & 2 \end{bmatrix} ,$$

$$X^{(4)} = \begin{bmatrix} 0 & 5 & 0 \\ 6 & 3 & 3 \end{bmatrix} , \quad X^{(5)} = \begin{bmatrix} 1 & 1 & 3 \\ 5 & 7 & 0 \end{bmatrix} , \quad X^{(6)} = \begin{bmatrix} 1 & 2 & 2 \\ 5 & 6 & 1 \end{bmatrix} .$$

Following Eqs. (3.11)–(3.13), the computations for the observed table $L = X^{(2)} = \begin{vmatrix} 0 & 3 & 2 \\ 6 & 5 & 1 \end{vmatrix}$ are

$$Q_L = \frac{5!\ 12!\ 6!\ 8!\ 3!}{17!} = 28{,}148.4163 \,,$$

$$R_{X^{(2)}} = 0!\ 3!\ 2!\ 6!\ 5!\ 1! = 1{,}036{,}800 \,,$$

and

$$P_{X^{(2)}} = \frac{28{,}148.4163}{1{,}036{,}800} = 0.0271 \ .$$

For reference, the exact probability values for the six 2 × 2 contingency tables listed above are:

$$P_{X^{(1)}} = \frac{28{,}148.4163}{6{,}220{,}800} = 0.0045 \ ,$$

$$P_{X^{(2)}} = \frac{28{,}148.4163}{1{,}036{,}800} = 0.0271 \ ,$$

$$P_{X^{(3)}} = \frac{28{,}148.4163}{829{,}440} = 0.0339 \ ,$$

$$P_{X^{(4)}} = \frac{28{,}148.4163}{3{,}110{,}400} = 0.0090 \ ,$$

$$P_{X^{(5)}} = \frac{28{,}148.4163}{3{,}628{,}800} = 0.0078 \ ,$$

and

$$P_{X^{(6)}} = \frac{28{,}148.4163}{345{,}600} = 0.0814 \ .$$

In an effort to expand Fisher's exact probability test to higher dimensions, Freeman and Halton examined three-dimensional contingency tables. A three-dimensional contingency table is more complex, but the approach by Freeman and Halton was an innovative permutation method. In the case, for example, of a 2 × 2 × 2 contingency table, there are

$$v = \prod_{m=1}^{k} r_m - \sum_{m=1}^{k} (r_m - 1) - 1 = (2)(2)(2) - (2-1) - (2-1) - (2-1) - 1 = 4$$

degrees of freedom; thus, only four cells need be manipulated with the remaining cells determined by the fixed marginal frequency totals. For Freeman and Halton, this meant the cell frequencies in the front 2 × 2 panel and the cell frequencies in the left uppermost cell of the rear 2 × 2 panel.[54]

[54] In many applications, "panels" are sometimes referred to as "slices" or "levels."

Consider the six 2 × 2 × 2 contingency tables listed here:

$$X^{(1)} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 3 \\ 0 & 5 & 3 & 14 \end{array}\right] , \qquad X^{(2)} = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 2 \\ 0 & 5 & 2 & 15 \end{array}\right] ,$$

$$X^{(3)} = \left[\begin{array}{cc|cc} 0 & 0 & 2 & 1 \\ 0 & 5 & 1 & 16 \end{array}\right] , \qquad X^{(4)} = \left[\begin{array}{cc|cc} 0 & 0 & 3 & 0 \\ 0 & 5 & 0 & 17 \end{array}\right] ,$$

$$X^{(5)} = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 3 \\ 1 & 4 & 2 & 15 \end{array}\right] , \qquad X^{(6)} = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 2 \\ 1 & 4 & 1 & 16 \end{array}\right] ,$$

where the vertical lines separate the front 2 × 2 panels from the rear 2 × 2 panels.

The expression for the probability of the tth three-dimensional contingency table as given by Freeman and Halton is

$$P_{X^{(t)}} = \frac{\text{[t]he product of all (border totals)!}}{(n!)^2 \times \text{ the product of all (cell totals)!}} . \tag{3.14}$$

More formally, given an $r \times c \times s$ contingency table with fixed marginal frequency totals a_i for $i = 1, \ldots, r$, b_j for $j = 1, \ldots, c$, and d_k for $k = 1, \ldots, s$, let n denote the total number of objects and let x_{ijk} denote the cell frequency for $i = 1, \ldots, r$, $j = 1, \ldots, c$, and $k = 1, \ldots, s$. As before, index each table by t. Then the probability of the $r \times c \times s$ contingency table is given by

$$P_{X^{(t)}} = \frac{\displaystyle\prod_{i=1}^{r} a_i! \prod_{j=1}^{c} b_j! \prod_{k=1}^{s} d_k!}{\displaystyle (n!)^2 \prod_{i=1}^{r} \prod_{j=1}^{c} \prod_{k=1}^{s} x_{ijk}^{(t)}!} .$$

Thus, for example,

$$P_{X^{(1)}} = \frac{5!\ 20!\ 3!\ 22!\ 3!\ 22!}{(25!)^2\ 0!\ 0!\ 0!\ 5!\ 0!\ 3!\ 3!\ 14!} = \frac{1{,}938}{13{,}225} = 0.1465 ,$$

where the panel marginals are $0 + 0 + 0 + 5 = 5$ and $0 + 3 + 3 + 14 = 20$, the row marginals are $0 + 0 + 0 + 3 = 3$ and $0 + 5 + 3 + 14 = 22$, and the column marginals are $0 + 0 + 0 + 3 = 3$ and $0 + 5 + 3 + 14 = 22$. Note that "[t]he product of all (border totals)!" in Eq. (3.14) and $n!$ are constants for all tables, thus "the product of all (cell totals)!" in Eq. (3.14) is operative here. For reference, the exact probability values for the six 2 × 2 × 2 contingency tables listed above are

$$P_{X^{(1)}} = \frac{5.5189 \times 10^{13}}{3.7661 \times 10^{14}} = 0.1465 ,$$

$$P_{X^{(2)}} = \frac{5.5189 \times 10^{13}}{6.2768 \times 10^{14}} = 0.0879 ,$$

$$P_{X^{(3)}} = \frac{5.5189 \times 10^{13}}{5.0215 \times 10^{15}} = 0.0110\ ,$$

$$P_{X^{(4)}} = \frac{5.5189 \times 10^{13}}{2.5609 \times 10^{17}} = 0.0002\ ,$$

$$P_{X^{(5)}} = \frac{5.5189 \times 10^{13}}{3.7661 \times 10^{14}} = 0.1465\ ,$$

and

$$P_{X^{(6)}} = \frac{5.5189 \times 10^{13}}{1.0043 \times 10^{15}} = 0.0550\ .$$

The recursive process described by Freeman and Halton simplified calculations and enabled computation of the probability value of a specified table from the probability value of a preceding table, provided the arrays were properly sequenced. Consider first contingency tables $X^{(1)}$ and $X^{(2)}$. Since the front panels of $X^{(1)}$ and $X^{(2)}$ are identical, it is only necessary to evaluate the rear panels of the two tables. Consider the ratio of the cell frequencies in the rear panel of $X^{(1)}$ to the cell frequencies in the rear panel of $X^{(2)}$; viz.,

$$P_{X^{(2)}} = P_{X^{(1)}} \times \frac{0!\ 3!\ 3!\ 14!}{1!\ 2!\ 2!\ 15!} = P_{X^{(1)}} \times \left[\frac{3 \times 3}{1 \times 15}\right] = 0.1465 \times \frac{9}{15} = 0.0879\ ,$$

which can easily be obtained from $X^{(1)}$ and $X^{(2)}$ as follows. For the rear panel in $X^{(1)}$ consider the two diagonal values in the upper-right and lower-left cells, e.g., 3 and 3, and for the rear panel in $X^{(2)}$ consider the two diagonal values in the upper-left and lower-right cells, e.g., 1 and 15, yielding the ratio in square brackets $\left[\frac{3\times3}{1\times15}\right]$. Next, consider the ratio of $X^{(2)}$ to $X^{(3)}$. Again, the front panels are identical, so the ratio of the cell values in the two rear panels is given by

$$P_{X^{(3)}} = P_{X^{(2)}} \times \frac{1!\ 2!\ 2!\ 15!}{2!\ 1!\ 1!\ 16!} = P_{X^{(2)}} \times \left[\frac{2 \times 2}{2 \times 16}\right] = 0.0879 \times \frac{4}{32} = 0.0110\ .$$

As Freeman and Halton noted, when considering the two ratios in square brackets, the ratio of $X^{(2)}$ to $X^{(3)}$ can be obtained from the preceding ratio of $X^{(1)}$ to $X^{(2)}$ by subtracting one from each value in the numerator (3×3), e.g., $3 - 1 = 2$ and $3 - 1 = 2$, and adding one to each value in the denominator (1×15), e.g., $1 + 1 = 2$ and $15 + 1 = 16$, thereby yielding $\left[\frac{2\times2}{2\times16}\right]$. Thus, to obtain the ratio of $X^{(3)}$ to $X^{(4)}$, subtract one from each value in the numerator (2×2), e.g., $2 - 1 = 1$ and $2-1 = 1$, and add one to each value in the denominator (2×16), e.g., $2+1 = 3$ and $16 + 1 = 17$, yielding $\left[\frac{1\times1}{3\times17}\right]$. Thus,

$$P_{X^{(4)}} = P_{X^{(3)}} \times \left[\frac{1 \times 1}{3 \times 17}\right] = 0.0110 \times \frac{1}{51} = 0.0002 \, .$$

At this point, the sequencing breaks down as there are two 1s in the numerator. However, as Freeman and Halton noted, $X^{(5)}$ can be obtained in alternative ways. For example,

$$P_{X^{(5)}} = P_{X^{(6)}} \times \left[\frac{1 \times 16}{2 \times 3}\right] = 0.0550 \times \frac{16}{6} = 0.1465 \, ,$$

where the numerator (1 × 16) is taken from the diagonal in the rear panel of $P_{X^{(6)}}$ and the denominator 2 × 3 is taken from the diagonal in the rear panel of $P_{X^{(5)}}$. Note that the front panels are identical in $P_{X^{(6)}}$ and $P_{X^{(5)}}$ and can therefore safely be ignored. Alternatively,

$$P_{X^{(5)}} = P_{X^{(1)}} \times \left[\frac{3 \times 5}{1 \times 15}\right] = 0.1465 \times \frac{15}{15} = 0.1465 \, ,$$

where the numerator (3×5) is taken from $P_{X^{(1)}}$ and the denominator (1×15) is taken from $P_{X^{(5)}}$. Here the front panels in $P_{X^{(1)}}$ and $P_{X^{(5)}}$ are different, thus the numerator and denominator values cannot be taken from only the rear panels of $P_{X^{(1)}}$ and $P_{X^{(5)}}$. The process is as follows with the eight cell frequency values of $P_{X^{(1)}}$ in the numerator and the eight cell frequency values of $P_{X^{(5)}}$ in the denominator:

$$\begin{aligned} P_{X^{(5)}} = P_{X^{(1)}} \times \frac{0!\;0!\;0!\;3!\;0!\;5!\;3!\;14!}{0!\;0!\;1!\;4!\;0!\;3!\;2!\;15!} \\ = P_{X^{(1)}} \times \left[\frac{3 \times 5}{1 \times 15}\right] = 0.1465 \times \frac{15}{15} = 0.1465 \, . \end{aligned}$$

Freeman and Halton concluded that the exact method they described was generally useful in cases where a chi-squared test would normally be utilized, but should not be used because the observed and expected cell frequencies were too small. The method, they explained, was also useful when a chi-squared test was wholly unsuitable, such as when the entire population contained so few members that a chi-squared test was not appropriate, but still a test of significance was required [480, p. 149]. Finally, they cautioned that a difficulty with the exact method described was the amount of labor involved in obtaining the exact probability values, thus setting an upper limit to the size of the sample that could be dealt with in a reasonable amount of time [480, p. 141].

3.24 Kruskal–Wallis and the C-sample Rank-Sum Test

In 1952 William Kruskal and W. Allen Wallis proposed an exact multiple-sample rank-sum test that they called H, and also provided tables for various levels of significance [779].

W.H. Kruskal

William Henry Kruskal earned his B.S. degree in mathematics and philosophy from Harvard University in 1940 and his M.S. degree in mathematics from Harvard University the following year. In 1941 Kruskal decided to take a job at the U.S. Naval Proving Ground in Dahlgren, Virginia. In 1946, Kruskal left the Navy and went to work in the family firm of Kruskal & Kruskal, a major fur wholesale business. In 1950, W. Allen Wallis offered Kruskal a position in the newly formed Department of Statistics at the University of Chicago, which he enthusiastically accepted [429, p. 257]. Kruskal went on to complete his Ph.D. in mathematical statistics from Columbia University in 1955. In addition to teaching at the University of Chicago, Kruskal also served as Chair of the Department of Statistics from 1966 to 1973, Dean of the Division of Social Sciences from 1974 to 1984, and Dean of the Irving B. Harris Graduate School of Public Policy Studies from 1988 to 1989 [429, 1484]. William Henry Kruskal died on 21 April 2005 in Chicago at age 85.

W.A. Wallis

Wilson Allen Wallis earned his Bachelor's degree in psychology from the University of Minnesota in 1932. He completed 1 year of graduate work at Minnesota, followed by a second year of graduate studies at the University of Chicago. In 1935 Wallis left the University of Chicago to study statistics under Harold Hotelling at Columbia University. As Wallis described it, "the only degree I ever got is a Bachelor's degree at Minnesota, except for the four honorary doctorates" [1056, p. 122]. From 1942 to 1946, Wallis was a member of the Statistical Research Group at Columbia, where he worked with such notables as Churchill Eisenhart, Milton Friedman, Fredrick Mosteller, Jimmy Savage, Herbert Solomon, George Stigler, Abraham Wald, and Jacob Wolfowitz, among others (q.v. page 69).

Wallis held faculty positions at Yale University, Stanford University, and the University of Chicago, and administrative positions at Columbia University, the University of Chicago, and the University of Rochester, where he was President from 1962 to 1970 and Chancellor from 1970 to 1982. Wallis served as the Under Secretary for Economic Affairs in the U.S. Department of State from 1982 to 1989. Wallis also served as an advisor to U.S. Presidents Dwight D. Eisenhower, Richard M. Nixon, Gerald R. Ford, and Ronald W. Reagan [814, 1056]. Wilson Allen Wallis died on 12 October 1998 in Rochester at the age of 85 [1410].

In 1952 Kruskal and Wallis introduced a C-sample rank-sum test statistic that they called H [779]. Although H is asymptotically distributed as chi-squared with $C - 1$ degrees of freedom, Kruskal and Wallis provided tables based on exact probability values for $C = 3$ with each sample less than or equal to 5 for $\alpha = 0.10$, 0.05, and 0.01 levels of significance.

Kruskal and Wallis explained that the H test statistic stems from two statistical methods: rank transformations of the original measurements and permutations of the rank-order statistics. They explained that if, in the one-way analysis of variance, the permutation method based on the conventional F statistic is combined with the rank method, the result is the H test.

Consider C random samples of possibly different sizes and denote the size of the ith sample by n_i, $i = 1, \ldots, C$. Let

$$N = \sum_{i=1}^{C} n_i$$

denote the total number of measurements, assign rank 1 to the smallest of the N measurements, rank 2 to the next smallest, and so on up to the largest measurement, which is assigned rank N, and let R_i denote the sum of the ranks in the ith sample, $i = 1, \ldots, C$. When there are no tied ranks, test statistic H is given by

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{C} \frac{R_i^2}{n_i} - 3(N+1) \; .$$

Kruskal and Wallis observed that when $C = 2$, H was equivalent to the Wilcoxon [1453], Festinger [427], Mann–Whitney [880], and Haldane–Smith two-sample rank-sum tests [573]. In 1953, in an erratum to their 1952 paper [779], Kruskal and Wallis documented the equivalence of the H test with the two-sample rank-sum test by van der Reyden that had recently come to their attention [1391]. In terms of permutation methods, Kruskal and Wallis provided a table of the distribution of H for $C = 3$ samples and sample sizes from one to five. They compared the exact probability values with three moment approximations, one based on the chi-squared distribution, one on the incomplete gamma distribution, and one on the incomplete beta distribution.

3.25 Box–Andersen and Permutation Theory

George Box and Sigurd Andersen read a paper on permutation tests and robust criteria before the Royal Statistical Society in November of 1954, which was subsequently published in *Journal of the Royal Statistical Society, Series B* in 1955 [193].

G.E.P. Box

George Edward Pelham Box, "Pel" to his friends, began college at the University of London as a chemistry student, but that work was interrupted by World War II when Box was called to military service. He served as a chemist in the British Army, but based on the work he was doing he quickly realized the importance of statistical training. Box had no background in statistics and, unable to find appropriate correspondence courses, taught himself the statistics he needed to conduct his work for the Army. Box returned to school after the War with a new interest in statistics, earning a B.Sc. degree in mathematical statistics from the University of London in 1947. Box worked for Imperial Chemical Industries (ICI) while he was completing his Ph.D. at the University of London, under the direction of Egon Pearson and H.O. Hartley. He earned his Ph.D. in statistics in 1952 and took leave from ICI to accept a visiting professorship at the Institute of Statistics, North Carolina State College (now, North Carolina State University at Raleigh), at the invitation of Gertrude Cox, 1953–1954.

In 1957, Box resigned from ICI to become Director of the Statistical Research Group at Princeton University. Two years later, Box married Joan G. Fisher, R.A. Fisher's daughter. In 1960, Box joined the faculty at University of Wisconsin at Madison to form a new department of statistics, where he remained for the rest of his career [338, 1418]. As Box noted in his memoirs, he was appointed to initiate and head a department of statistics as a full professor even though he had never had an academic appointment at any university [192, p. 95]. Box was elected Fellow of the Royal Society in 1985 and in 1992 Box retired from the University of Wisconsin. (Several sources report the year in which Box received his F.R.S. as 1979, but 21 March 1985 appears to be correct; see [192, p. 245].)

An interesting aside: Box is credited with coining the term "robustness" in a 1953 article that appeared in *Biometrika* on non-normality and tests on variances [190, p. 318]. However, John Hunter reported that Box had remarked in his acceptance letter to Gertrude Cox that in addition to research on the design of experiments he hoped to look into the problem of robust statistics. Hunter stated "I believe that this is the first time the word 'robust' appears in a statistics context" [998]. George Box published over 200 journal articles in his lifetime, the first at age 19 [192, pp. xviii, 17]. George Edward Pelham Box F.R.S. died at home on 28 March 2013 in Madison, Wisconsin, at age 93. Just days before his death, advance copies of his autobiography were flown out to him by his publisher John Wiley & Sons; interested readers should consult *An Accidental Statistician: The Life and Memories of George E. P. Box* [192].

S.L. Andersen

Sigurd Lökken Andersen was born in Silkeborg, Denmark, and at the age of three, emigrated by ship to the United States with his parents. Andersen began his education at Princeton University, but left when World War II began to enlist in the Navy and later continued his education at Cornell University. Upon graduation from Cornell, Anderson served at sea in the North Atlantic. At the conclusion of World War II, Andersen enrolled at North Carolina State University at Raleigh where he was assigned as a research assistant to George Box, along with John Stuart (Stu) Hunter. [338, 998]. Andersen received his Ph.D. from North Carolina State University at Raleigh in 1954 with a dissertation on robust tests for variances under the direction of Robert John Hader. Possibly the reason why Hader is listed as Andersen's dissertation advisor instead of Box is because the research was funded by the Office of Ordnance Research, United States Army, under contract DA-36-034-ORD-1177, which was administered by Hader. After graduation, Andersen took a position with the DuPont Corporation in Wilmington, Delaware, remaining there for 35 years until his retirement in 1989 [874]. Sigurd Lökken Andersen died on 5 August 2012 at the age of 88.

In November of 1954 Box and Andersen read a paper on "Permutation theory in the derivation of robust criteria and the study of departures from assumption" before the Research Section of the Royal Statistical Society, subsequently published under the same title in *Journal of the Royal Statistical Society, Series B* in 1955 [193]. This is a lengthy paper and includes discussions by several members of the Society. Unfortunately, the sheer length of the paper precludes anything but the briefest summary and it is not possible to do justice to this important paper in this limited space. Box and Andersen noted that in practical circumstances little is usually known of the validity of assumptions, such as the normality of the error distribution. They argued for statistical procedures that were insensitive to changes in extraneous factors not under test, but sensitive to those factors under test, i.e., procedures both robust and powerful. In this context, they addressed permutation theory as a robust method and applied it to comparisons of means and variances.

It is important to note that Box and Andersen found most of the standard normal-theory tests to compare means to be "remarkably robust" and sufficient to fulfill the needs of researchers. As they explained, "our object in discussing permutation theory for these tests is to demonstrate this [robustness], and to consider more clearly the behavior of the permutation tests in those cases with which we are most familiar" [193, p. 33]. They emphasized, however, that their object was not to suggest alternative tests for these research situations.

Box and Andersen pointed out that tests on differences between variances could be so misleading as to be valueless, unless the resulting distribution was very close to normal. They then stated "[t]he authors' belief is that such an assertion [of normality] would certainly not be justified" [193, p. 2]. The solution, they concluded, was in the use of "a remarkable new class of tests" called permutation tests, such as introduced by R.A. Fisher in 1935 [451]. Box and Andersen distinguished between two alternative views of the nature of inference in permutation tests. In the first view, a data-dependent inference was confined only to that finite population of samples produced by rearrangement of the observations of the experiment. In the second view that was not a data-dependent inference, the samples were to be regarded as being drawn from some hypothetical infinite population in the usual way. It was the second alternative that was preferred by Box and Andersen.

Like others in this era, Box and Andersen observed that evaluation of a permutation distribution is laborious and in order to make permutation theory of practical value, researchers such as Pitman and Welch used an approximation to the permutation distribution based on the value of its moments, e.g., the beta distribution [1129, 1130, 1428, 1430]. After defining a modified F test where the degrees of freedom were adjusted to compensate for non-normality and differences among variances, they considered two questions:

1. How good is the moment approximation to the permutation test?
2. How much power is lost by using the modified F test when the distribution happens to be normal?

They then investigated the power and robustness of the standard F test and the modified F test for the rectangular, normal, and double-exponential parent distributions.

In the conclusion to the paper they noted that one of the simplest statistical procedures was the test of hypothesis that the mean of a sample, $\bar{x}$, is equal to the mean of the population, μ_x, when the population standard deviation, σ_x, was known. They explained that if a sample of n observations $\{x_1, x_2, \ldots, x_n\}$ was available, the criterion usually chosen was $\sqrt{n}\,\bar{x}/\sigma_x$, which was then referred to tables of the unit normal distribution. Box and Andersen noted that the validity of this test of the null hypothesis does not depend on the supposition that the observations are exactly normally distributed, as the central limit theorem guarantees that, for almost all parent distributions, the chosen statistic is asymptotically distributed in the assumed form. They concluded that "for all but extremely small sample sizes and 'pathological' parent distributions the null test is approximately valid" [193, p. 25]. They further noted that a similar argument may be employed to analysis of variance tests. However, if the analysis of variance lacks the central limit property, it is necessary to seek alternative tests with greater robustness. One way of doing this is by approximating to the appropriate permutation test. Thus, for Box and Andersen the permutation test was implicitly treated as a gold standard against which the F test was to be evaluated.

Fig. 3.12 A 2 × 2 contingency table in the notation of Leslie [821]

x	$n_A - x$	n_A
$n_B - x$	$N - n_B - n_A + x$	$N - n_A$
n_B	$N - n_B$	N

3.26 Leslie and Small Contingency Tables

In 1955 Patrick Leslie proposed a new method for calculating the exact probability value of a 2 × 2 contingency table that was based on ordinary binomial coefficients, which could easily be obtained from Pascal's triangle [821].

P.H. Leslie

Patrick Holt Leslie, known to his family and friends as "George," was educated at Westminster School and Christ Church College, University of Oxford, where he obtained an honors degree in physiology in 1921, but was prevented from pursuing a medical degree due to a serious lung disease. After several years of research in bacteriology in the School of Pathology at the University of Oxford, his remarkable flair for mathematics came to be recognized and at age 35 he turned to statistical theory and population dynamics with the Bureau of Animal Population. He continued that work from 1935 until his retirement from the Bureau in 1967. Later in life, Leslie received a D.Sc. from the University of Oxford based on the published results of his various research projects. Born with the century, Patrick Holt Leslie died in June 1972 at the age of 72 [23, 43, 298, p. 18].

In 1955 Leslie published a short paper of only one-and-a-half folio pages on "a simple method of calculating the exact probability in 2 × 2 contingency tables with small marginal totals" [821]. A 2 × 2 contingency table in Leslie's notation is given in Fig. 3.12 where N denotes the total number of observations and the marginal frequency totals fulfill

$$n_A \leq N - n_A, \; n_B \leq N - n_B, \; n_B \leq n_A, \; a_x = \binom{n_A}{x}, \text{ and } b_x = \binom{N - n_A}{n_B - x} \; ;$$

then

$$C = \sum_{x=0}^{n_B} a_x b_x = \frac{N!}{n_B! \, (N - n_B)!} \; .$$

Calculation of the exact hypergeometric one-tailed cumulative probability for the appropriate tail of the distribution is then easily obtained from a_x and

Fig. 3.13 Example 2 × 2 contingency table with $N = 16$ cases

5	2	7
1	8	9
6	10	16

Fig. 3.14 Binomial coefficients for $n_A = 7$ with a_x, $x = 0, \ldots, n_B$ and b_x, $x = n_B, \ldots, 0$

x	a_x	b_x	$a_x b_x$
0	1	84	84
1	7	126	882
2	21	126	2,646
3	35	84	2,940
4	35	36	1,260
5	21	9	189
6	7	1	7

b_x, $x = 0, \ldots, n_B$. Once the appropriate tail is determined, the sum of the $a_x b_x$ products for $w \leq x$ for the left tail or $w \geq x$ for the right tail are

$$\frac{1}{C} \sum_{w=0}^{x} a_w b_w \text{ for the left tail }, \quad \text{or} \quad \frac{1}{C} \sum_{w=x}^{n_B} a_w b_w \text{ for the right tail },$$

yielding a one-tailed exact cumulative probability value.

Consider a simple example with $n_A = 7$, $n_B = 6$, $N = 16$, and $x = 5$; the completed 2 × 2 contingency table is shown in Fig. 3.13. The essential values are the binomial coefficients for $n_A = 7$, constituting a_x, $x = 0, \ldots, n_B$, and in reverse order the binomial coefficients for $N - n_A = 9$, constituting b_x, $x = n_B, \ldots, 0$, as given in Fig. 3.14. The required binomial coefficients can easily be obtained from the first $n + 1$ terms of the expanded binomial series,

$$1 + \frac{n}{1!} + \frac{n(n-1)}{2!} + \frac{n(n-1)(n-2)}{3!} + \cdots + \frac{n!}{n!} = \sum_{i=0}^{n} \binom{n}{i} = 2^n \,.$$

Also, the required binomial coefficients can be obtained by enumerating Pascal's triangle up to the required marginal frequency total. For reference, Table 3.11 displays Pascal's triangle containing the requisite binomial coefficients for $n = n_B = 7$ and $n = N - n_A = 9$. To be faithful to Pascal, Fig. 3.15 shows Pascal's arithmetical triangle as he actually laid it out in *Traité du triangle arithmétique* in 1665 [399, Frontispiece].[55]

[55] As noted by Edwards [399, p. x], it was Pierre Raymond de Montmort who, in 1708, first attached the name of Pascal to the combinatorial triangle; however, he changed the form to a staggered version [334]. Then, in his *Miscellanea Analytica* of 1730, Abraham de Moivre christened Pascal's original triangle "Triangulum Arithmeticum PASCALIANUM."

Table 3.11 Pascal's triangle for $n = 0, \ldots, 9$

	Binomial coefficients									
n	1	2	3	4	5	6	7	8	9	10
0	1									
1	1	1								
2	1	2	1							
3	1	3	3	1						
4	1	4	6	4	1					
5	1	5	10	10	5	1				
6	1	6	15	20	15	6	1			
7	1	7	21	35	35	21	7	1		
8	1	8	28	56	70	56	28	8	1	
9	1	9	36	84	126	126	84	36	9	1

Fig. 3.15 Pascal's arithmetical triangle as originally published in 1665 [399, Frontispiece]

Pascal's triangle

1	1	1	1	1	1
1	2	3	4	5	
1	3	6	10		
1	4	10			
1	5				
1					

The sum of the $a_x b_x$ product column in Fig. 3.14 is

$$C = \sum_{x=0}^{n_B} a_x b_x = 84 + 882 + \cdots + 7 = 8{,}008 \ ,$$

and since the observed cell frequency of $x = 5$ in the cell with the smallest expectation is greater than the expectation given by $(7 \times 6)/16 = 2.625$, $x = 5$ lies in the right tail of the distribution. Thus $a_x b_x$ for $n = 5$ and $n = 6$, is $189+7 = 196$ and the one-tailed exact cumulative probability value is

$$\frac{1}{C} \sum_{w=x}^{n_b} a_w b_w = \frac{1}{8{,}008} 196 = 0.0245 \ .$$

3.27 A Two-Sample Rank Test for Dispersion

Wilcoxon in 1945, Festinger in 1946, Mann and Whitney in 1947, Haldane and Smith in 1948, and van der Reyden in 1952 had developed two-sample rank-sum tests wherein the sum of the ranks of one of the samples was used in a test of hypothesis that the two samples came from the same population, i.e., the hypothesis of the equivalence of the two distribution functions [427,880,1391,1441, 1453]. Such tests are sensitive to possible differences in location between the two distribution functions. In the 1950s, two non-parametric tests were published for the equivalence of parameters of dispersion, both assuming that the location parameters were equal and both yielding tables of probability values. The first of the published papers was by Sidney Rosenbaum in 1953 and the second was by Anant Kamat in 1956.

3.27.1 Rosenbaum's Rank Test for Dispersion

In 1953 S. Rosenbaum proposed a rank test for the equivalence of parameters of dispersion, assuming that the location parameters (mean or median) were equal [1193].

S. Rosenbaum

Sidney Rosenbaum was born in London in 1918 and educated at the University of Cambridge. Rosenbaum received an emergency commission in the Royal Regiment of Artillery in 1943 and 2 years later became a temporary Captain. From 1951 to 1963 Rosenbaum served as the Principal Scientific Officer, Army Medical Statistics Branch, War Office, and during that time he also worked on his doctorate, earning his Ph.D. in Medical Statistics at the London School of Hygiene & Tropical Medicine in 1960 [615]. He became the Chief Statistician at the Department of the Treasury and transferred to the Cabinet Office before his final appointment as Director of Statistics and Operation Research at the Civil Service College in 1972 [24]. After retirement, Rosenbaum worked as a consultant to the Department of Health and Social Security (DHSS) and the Ministry of Technology. Rosenbaum was one of the longest-serving Fellows of the Royal Statistical Society, having been elected in 1948. He was elected a Fellow of the Royal Society of Medicine in 1956. Sidney Rosenbaum passed away in March of 2013 at the age of 94 [24].

To illustrate the Rosenbaum rank test for dispersion, consider a sample of n points and a second sample of m points from a population with a continuous distribution function. As Rosenbaum noted, the probability that r points of the sample of m will lie outside the end values of the sample of n is given by

$$\begin{aligned} P_r &= n(n-1)\frac{m!}{(m-r)!} \times \frac{(r+1)(n+m-r-2)!}{(n+m)!} \\ &= n(n-1)\binom{m}{r} B(n+m-1-r,\, r+2)\ , \end{aligned}$$

where B is the complete beta function.

Then, for $r_0 \leq m$,

$$\sum_{r=0}^{r_0} P_r = \sum_{r=0}^{r_0} n(n-1)\binom{m}{r} B(n+m-1-r,\, r+2)$$

is the probability that the value of r is not greater than r_0. As was usual during this time, traditional fixed levels of significance took precedence over exact probability values. Thus, Rosenbaum fixed a probability level ε and arrived at an r_0 such that

$$\sum_{r=0}^{r_0-1} P_r \leq \varepsilon < \sum_{r=0}^{r_0} P_r \ .$$

Rosenbaum provided tables of $r = r_0 + 1$ for $\varepsilon = 0.95$ and 0.99 over the range $n = 2, \ldots, 50$ and $m = 2, \ldots, 50$. The tables give the probability values, less than 0.05 and 0.01, that r or more points of a sample of size m lie outside the extreme values of a sample of size n if the samples are drawn from the same population, whatever its distribution [1193, pp. 665, 667].

3.27.2 Kamat's Rank Test for Dispersion

Motivated by Rosenbaum's 1953 article, in 1956 A.R. Kamat proposed an alternative rank test for the equivalence of parameters of dispersion, assuming that the location parameters were equal [707].

A.R. Kamat

Anant Raoji Kamat was born in 1912 and completed his early education in the Ratnagiri district of Maharashtra. Kamat was an exceptional student, entering the University of Bombay (Mumbai) in 1929 where he received his undergraduate and M.Sc. degrees. In 1953, Kamat earned his Ph.D. in mathematical statistics at the University of London. Kamat was drawn to political activism, but also worked as an academic, teaching courses in mathematics and statistics. A gifted social scientist, Kamat joined the Gokhale Institute of Politics and Economics, University of Poona (Pune), in 1959, retiring from there as Joint Director in 1974. After his retirement, Kamat received a fellowship to the Indian Council of Social Science Research to continue to work on issues of education. Anant Raoji Kamat passed away on 9 July 1983 at the age of 71 [1019].

To illustrate Kamat's rank test for dispersion, consider two samples x_i, $i = 1, \ldots, n$, and y_j, $j = 1, \ldots, m$, with $m \geq n$. Pool and rank the measurements in order and let R_n and R_m denote the range of ranks of x and y, respectively. The test statistic proposed by Kamat was

$$D_{n,m} = R_n - R_m + m ,$$

where $D_{n,m}$ can take values $0, 1, \ldots, m + n$. For example, if $n = 4$ with ranks $\{2, 4, 6, 8\}$ and $m = 5$ with ranks $\{1, 3, 5, 7, 9\}$, then $R_n = R_4 = 8 - 2 = 6$, $R_m = R_5 = 9 - 1 = 8$, and $D_{n,m} = D_{4,5} = 6 - 8 + 5 = 3$. Large and small values of $D_{n,m}$ indicate possible divergence from the hypothesis that the parameters of dispersion of the populations from which the samples were drawn are equal [707, p. 377]. Kamat provided a table of percentage points of $D_{n,m}$ based on exact probability values for $m + n \le 20$ with $\alpha = 0.05, 0.025, 0.005$, and 0.001.

The technique used by Kamat to generate the permutation distribution of $D_{n,m}$ was based on simple combinatorial rules and is worth explaining in some detail. The total number of ways that the $m + n$ ranks can be arranged is given by

$$\binom{m+n}{n} ,$$

thus providing the total number of values of $D_{n,m}$. However, as Kamat explained, the total number of ways can be constructed in another manner, from four separate procedures:

(a) $R_n = n - 1$; $R_m = m - 1$; $D_{n,m} = n$.
This result can be achieved in only two ways.

(b) $R_n = n - 1 + i,\ i = 0, \ldots, m - 1$; $R_m = m - 1 + n$; $D_{n,m} = i$.
This result can be achieved in

$$(m - 1 - i)\binom{n + i - 2}{n - 2} \text{ ways} . \tag{3.15}$$

(c) $R_n = n - 1 + m$; $R_m = m - 1 + j,\ j = 0, \ldots, n - 1$; $D_{n,m} = n - m - j$.
By symmetry, this result can be achieved in

$$(n - 1 - j)\binom{m + j - 2}{m - 2} \text{ ways} . \tag{3.16}$$

(d) $R_n = n - 1 + i,\ i = 0, \ldots, m - 1$; $R_m = m - 1 + j,\ j = 0, \ldots, n - 1$; $D_{n,m} = n + i - j$.
This result can be achieved in

$$2\binom{i + j - 2}{j - 1} \text{ ways} , \tag{3.17}$$

where $i = 1, \ldots, m - 1$ and $j = 1, \ldots, n - 1$.

Table 3.12 Combinations of $m + n = 4 + 3 = 7$ ranks considered $n = 3$ at a time with associated R_n, R_m, and $D_{n,m}$ values

	Sequence				
Number	$n = 3$	$m = 4$	R_3	R_4	$D_{3,4}$
1	1 2 3	4 5 6 7	2	3	3
2	5 6 7	1 2 3 4	2	3	3
3	2 3 4	1 5 6 7	2	6	0
4	2 3 5	1 4 6 7	3	6	1
5	2 4 5	1 3 6 7	3	6	1
6	3 4 5	1 2 6 7	2	6	0
7	2 3 6	1 4 5 7	4	6	2
8	2 4 6	1 3 5 7	4	6	2
9	3 4 6	1 2 5 7	3	6	1
10	2 5 6	1 3 4 7	4	6	2
11	3 5 6	1 2 4 7	3	6	1
12	4 5 6	1 2 3 7	2	6	0
13	1 2 7	3 4 5 6	6	3	7
14	1 3 7	2 4 5 6	6	4	6
15	1 4 7	2 3 5 6	6	4	6
16	1 5 7	2 3 4 6	6	4	6
17	1 6 7	2 3 4 5	6	3	7
18	1 2 4	3 5 6 7	3	4	3
19	1 3 4	2 5 6 7	3	5	2
20	1 2 5	3 4 6 7	4	4	4
21	1 3 5	2 4 6 7	4	5	3
22	1 4 5	2 3 6 7	4	5	3
23	1 2 6	3 4 5 7	5	4	5
24	1 3 6	2 4 5 7	5	5	4
25	1 4 6	2 3 5 7	5	5	4
26	1 5 6	2 3 4 7	5	5	4
27	2 3 7	1 4 5 6	5	5	4
28	2 4 7	1 3 5 6	5	5	4
29	3 4 7	1 2 5 6	4	5	3
30	2 5 7	1 3 4 6	5	5	4
31	3 5 7	1 2 4 6	4	5	3
32	4 5 7	1 2 3 6	3	5	2
33	2 6 7	1 3 4 5	5	4	5
34	3 6 7	1 2 4 5	4	4	4
35	4 6 7	1 2 3 5	3	4	3

To illustrate Kamat's technique, consider $n = 3$ and $m = 4$ measurements, pooled and ranked from 1 to $m + n = 4 + 3 = 7$. Table 3.12 lists the 35 possible sequences, divided into $n = 3$ and $m = 4$ ranks, with values for R_3, R_4, and $D_{3,4}$. Sequences 1 and 2 in Table 3.12 are the two ways possible under (a).

Sequences 3 through 12 in Table 3.12 are the ten ways possible under (b), i.e., following Eq. (3.15) for $i = 0$,

$$(4-1-0)\binom{3+0-2}{3-2} = (3)\binom{1}{1} = 3 \text{ ways ;}$$

for $i = 1$,

$$(4-1-1)\binom{3+1-2}{3-2} = (2)\binom{2}{1} = 4 \text{ ways ;}$$

for $i = 2$,

$$(4-1-2)\binom{3+2-2}{3-2} = (1)\binom{3}{1} = 3 \text{ ways ;}$$

and for $i = 3$,

$$(4-1-3)\binom{3+3-2}{3-2} = (0)\binom{4}{1} = 0 \text{ ways .}$$

Sequences 13 through 17 in Table 3.12 are the five ways possible under (c), i.e., following Eq. (3.16) for $j = 0$,

$$(3-1-0)\binom{4+0-2}{4-2} = (2)\binom{2}{2} = 2 \text{ ways ;}$$

for $j = 1$,

$$(3-1-1)\binom{4+1-2}{4-2} = (1)\binom{3}{2} = 3 \text{ ways ;}$$

and for $j = 2$,

$$(3-1-2)\binom{4+2-2}{4-2} = (0)\binom{4}{2} = 0 \text{ ways .}$$

Sequences 18 through 35 in Table 3.12 are the 18 ways possible under (d), i.e., following Eq. (3.17) for $i = 1$ and $j = 1$,

$$(2)\binom{1+1-2}{1-1} = (2)\binom{0}{0} = 2 \text{ ways ;}$$

for $i = 1$ and $j = 2$,

$$(2)\binom{1+2-2}{2-1} = (2)\binom{1}{1} = 2 \text{ ways ;}$$

for $i = 2$ and $j = 1$,

$$(2)\binom{2+1-2}{1-1} = (2)\binom{1}{0} = 2 \text{ ways} ;$$

for $i = 2$ and $j = 2$,

$$(2)\binom{2+2-2}{2-1} = (2)\binom{2}{1} = 4 \text{ ways} ;$$

for $i = 3$ and $j = 1$,

$$(2)\binom{3+2-2}{1-1} = (2)\binom{2}{0} = 2 \text{ ways} ;$$

and for $i = 3$ and $j = 2$,

$$(2)\binom{3+2-2}{2-1} = (2)\binom{3}{1} = 6 \text{ ways} .$$

Kamat then showed that by combining these four cases, (a)–(d), the probability of $D_{n,m}$ would be given by

$$P\{D_{n,m}\} = \frac{1}{\binom{m+n}{n}} \left\{ 2A_r \sum_{j=1}^{m} \binom{r-n+2j-2}{r-n+j-1} + 2B_r \sum_{i=1}^{n} \binom{n-r+2i-2}{n-r+i-1} + C_r(m-1-r)\binom{n+r-2}{n-2} + D_r(r-m-1)\binom{2m+n-r-2}{m-2} + 2E_r \right\} ,$$

where

$$A_r = \begin{cases} 1 & \text{if } r \leq m , \\ 0 & \text{otherwise} , \end{cases} \quad B_r = \begin{cases} 1 & \text{if } r > m , \\ 0 & \text{otherwise} , \end{cases} \quad C_r = \begin{cases} 1 & \text{if } r < m , \\ 0 & \text{otherwise} , \end{cases}$$

$$D_r = \begin{cases} 1 & \text{if } r > m , \\ 0 & \text{otherwise} , \end{cases} \quad E_r = \begin{cases} 1 & \text{if } r = m , \\ 0 & \text{otherwise} , \end{cases}$$

and $r = n + i - j$.

Exhibiting some frustration with calculation difficulties in the absence of high-speed computers, Kamat noted that calculation of percentage points from the exact distribution becomes impractical when the sequence becomes large [707, p. 379]. The remainder of Kamat's article was devoted to finding a suitable approximation based on the first three moments of $D_{n,m}$. Finally, Kamat noted that when one sample (say, the m sample) is wholly included within the extreme values of the other sample, then the Rosenbaum test statistic r and the Kamat test statistic $D_{n,m}$ are connected by the relation $D_{n,m} = m + r$.

3.28 Dwass and Modified Randomization Tests

Meyer Dwass is often credited with introducing resampling procedures for permutation tests, which he termed "modified permutation tests" [1431].

M. Dwass

Meyer Dwass earned his B.A. degree in mathematics from George Washington University in 1948, his M.A. degree in mathematical statistics from Columbia University in 1949, and his Ph.D. in statistics from the University of North Carolina at Chapel Hill in 1952 under Wassily Hoeffding. Dwass immediately took a position as Assistant Professor of Mathematics at Northwestern University where he remained for the rest of his academic career, with the exception of a brief time spent at the University of Minnesota from 1961 to 1962. Dwass was Chair of the Department of Mathematics at Northwestern from 1978 to 1981 and established the Department of Statistics at Northwestern in 1986. Meyer Dwass retired from Northwestern in 1989 and passed away on 15 July 1996 at the age of 73 [562, 1485].

While researchers prior to Dwass certainly utilized resampling to provide approximate probability values, such as Eden and Yates in their 1933 investigation into height measurements of Yeoman II wheat shoots in which they drew a sample of 1,000 out of a possible 4,586,471,424 permutations (q.v. page 39) [379], Dwass provided the first rigorous investigation of the precision of resampling probability approximations. In 1957 Dwass published an article on modified randomization tests for non-parametric hypotheses [368], which relied heavily on the theoretical contributions of Lehmann and Stein's 1949 article [818]. Dwass noted that a practical shortcoming of exact permutation procedures was the great difficulty in enumerating all the possible arrangements of the observed data. To illustrate, consider as Dwass did, two samples of sizes m and n. Dwass observed that even after elimination of those permutations yielding the same value of the statistic, the number of permutations could still be prohibitively large.[56] Thus, for sample sizes $m = n = 5$, there are

$$(m + n)! = (5 + 5)! = 10! = 3{,}628{,}800$$

permutations of the observed data to be considered, but only

$$\binom{m+n}{m} = \binom{5+5}{5} = \frac{10!}{5!\,5!} = 252$$

combinations of the observed data to be examined. However, for sample sizes as small as $m = n = 10$, there are still

$$\binom{m+n}{m} = \binom{10+10}{10} = \frac{20!}{10!\,10!} = 184{,}756$$

combinations of the observed data to be examined.

Dwass then proposed "the most obvious procedure" of examining a random sample drawn without replacement from all possible permutations and "making the decision to accept or reject the null hypothesis on the basis of those permutations only" [368, p. 182], as suggested by Eden and Yates much earlier [379]. Dwass determined bounds for the ratio of the power of the original procedure, in this case a two-sample test, to the resampling procedure and provided a table containing numerical values of the bounds. Note that in this table Dwass did not compare bounds from exact and resampling permutation procedures, but unfortunately compared bounds from a resampling probability procedure with those from a normal distribution. Letting s denote the number of resamplings, Dwass made computations for only those values of s such that $\alpha(s + 1)$ was an integer. Thus the table provided bounds for the ratio of the power of a two-sample test with a resampling test for values of $s = 19, 39, 49, 59, 79, 99, 119, 149, 199, 299, 499$, and 999 and for $\alpha = 0.01, 0.02, 0.05$, and 0.10. Examination of the table by Dwass reveals reasonably close agreement between the resampling approximate probability values and the approximate probability values obtained from a normal distribution. For example, let $s = 99$, then for $\alpha = 0.01$ the resampling and normal approximate bounds are 0.634 and 0.618, respectively; for $\alpha = 0.02$ the resampling and normal approximate bounds values are 0.732 and 0.726, respectively; for $\alpha = 0.05$ the resampling and normal approximate bounds are 0.829 and 0.827, respectively; and for $\alpha = 0.10$ the resampling and normal approximate bounds are both 0.881 and 0.881 [368, p. 182].

[56] As Box and Andersen noted in 1955, although there are $(m + n)!$ possible arrangements of a sample, there are only $(m + n)!/(m!\,n!)$ arrangements that result in possibly different mean differences [193, p. 7].

The main point made by Dwass was that instead of basing a statistical decision on all possible permutations of the observations, the statistical test could be based instead on a smaller number of permutations randomly selected from the set of all permutations and the power of the test would be "close" to that of the most powerful non-parametric test. Dwass observed that while it is true that s has to be very large, the optimum exact test is usually completely impossible. He posited that if $m = n = 20$, then

$$\binom{m+n}{m} = \binom{20+20}{20} = \frac{40!}{20!\ 20!} > 10^{11}$$

(actually, 137,846,528,820) and if a machine existed that could check 10 permutations per second, the job would run something on the order of 1,000 years [368, p. 185].[57]

Monte Carlo Methods

Stanislaw Marcin Ulam, Polish refugee and celebrated mathematician who worked on the Manhattan Project at the Los Alamos National Laboratory in Los Alamos, New Mexico, spent hours playing games of Canfield solitaire while recuperating from encephalitis in 1946. In so doing, he speculated about the odds of any randomly dealt hand. He filled page after page with probabilistic equations, but the problem proved intractable and he decided it was better to play a hundred random hands and tabulate what percentage of the time he won [372]. Unlike an experiment, the results were not certain, but the probability was sure to be very close. In later years, Ulam explained that the approach was named "Monte Carlo" in memory of an uncle who liked to gamble on the "well-known generator of random integers . . . in the Mediterranean principality [of Monte Carlo]" [712, pp. 109–111]. The term "Monte Carlo method" was coined in 1946 by Ulam, John von Neumann, and Nicholas Metropolis while they were working on nuclear weapons projects at the Los Alamos National Laboratory [927, 1419]. However, George Dyson attributes the coining of the term "Monte Carlo" to Nicholas Metropolis [370, p. 192].

The Monte Carlo method was quickly brought to bear on problems pertaining to thermonuclear as well as fission devices, and in 1948 Ulam reported to the Atomic Energy Commission about the application of the

(continued)

[57] Presently, resampling permutation routines, which are essentially sampling without replacement routines, generate hundreds of thousands of permutations per second when powered by an efficient uniform pseudorandom number generator (PRNG) such as the Mersenne Twister (MT) or the SIMD-oriented Fast Mersenne Twister (SFMT) on high-speed work stations [905, 1214].

Monte Carlo method for such things as cosmic ray showers and the study of the Hamilton Jacobi partial differential equation [372]. By 1949, applications of the Monte Carlo method discussed in the literature were many and varied and in that year a symposium on the Monte Carlo method—sponsored by the RAND Corporation, the National Bureau of Standards' Institute for Numerical Analysis, and the Oak Ridge Laboratory—was held at the University of California, Los Angeles [370, p. 198]. Later, a second symposium was organized by members of the Statistical Laboratory at the University of Florida in Gainesville [654, 926]. By 1987 it was reported that 10 billion uniform pseudorandom numbers were being generated on computers around the world for Monte Carlo solutions to problems that Ulam first dreamed about 40 years previously [359].

While Dwass is usually credited with the introduction of Monte Carlo resampling procedures for permutation tests, he was not the first to develop such procedures. Today, Monte Carlo methods in physics are used in the design of nuclear reactors, criticality analysis, oil well logging, health-physics problems, determinations of radiation doses, spacecraft radiation modeling, radiation damage studies, and research on magnetic fusion [359]. In addition, Monte Carlo methods are popular in statistics, economics, chemistry, astronomy, engineering, and even stock market analysis. For an extensive survey of Monte Carlo methods, including a bibliography of some 251 references, see a 1970 article on "A retrospective and prospective survey of the Monte Carlo method" by John Halton in *SIAM Review* [579].

Finally, in the context of the rank-order tests so common in the 1940s and 1950s, and on which Dwass did his dissertation, Dwass posed the following question: "For what value of s is the modified [resampling] test already better than some given *rank order* test, or in particular, than the rank order test which is best against the alternative under consideration?"[58] [368, p. 185].

3.29 Looking Ahead

Permutation methods are by their very nature computationally-intensive and permutation methods in the period between 1940 and 1959 were characterized by researchers expressing frustration over difficulties in computing a sufficient number of permutations of the observed data in a reasonable time. To compensate for the difficulty, many researchers turned to rank-order statistics, which were much more amenable to permutation methods. Thus, this period was distinguished by a plethora of rank-order tests. Examples included the Kendall rank-order correlation coefficient [728, 734], the Friedman two-way analysis of variance for ranks [485,

[58] Emphasis in the original.

486], the Wilcoxon two-sample rank-sum test [1453], the Festinger two-sample rank-sum test [427], the Mann–Whitney two sample rank-sum test [880], and the Kruskal–Wallis one-way analysis of variance rank test [779]. This led ipso facto to the publication of numerous tables of exact probability values for rank-order tests. Examples include tables for testing randomness by Swed and Eisenhart [1337]; for 2 × 2 contingency tables by Finney [434]; for the Spearman rank-order correlation coefficient by David, Kendall, and Stuart [328]; for the Wilcoxon–Mann–Whitney two-sample rank-sum test by Wilcoxon [1453, 1454], White [1441], and Fix and Hodges [465]; and for the two-sample rank-sum Mann–Whitney statistic by van der Reyden [1391] and Auble [40]. The end of the period saw an emphasis on the power of permutation tests compared with their conventional parametric counterparts by Hoeffding [636], Silvey [1275, 1276], and Box and Andersen [193], and the formal introduction of resampling techniques by Dwass in 1957 [368].

The development of computing continued unabated in the 1960s and 1970s with increases in memory, speed, and availability to researchers. New computer programming languages, interpreters, and operating systems were released in this period and the personal computer became generally available. The advent of accessible and efficient computers by researchers in the 1960s meant that the next two decades witnessed a proliferation of computer routines and algorithms designed to generate all permutation sequences of observed data sets, random permutation sequences of observed data sets, and permutations of cell frequencies in contingency tables. Many of these routines were designed by computer scientists and not by statisticians, but statisticians applied them to statistical problems such as permutation versions of paired and unpaired t and F tests, as well as various analyses of cross-classification contingency tables.

In addition, this period saw the introduction of a number of permutation tests, including the Siegel–Tukey test for relative spread in 1960 [1273]; the Mielke–Siddiqui matrix occupancy test in 1965 [988]; the Baker–Collier analysis of variance F test in 1966 [51, 52]; the Fisher–Yates exact probability test by Ghent in 1972 [510]; multi-response permutation procedures by Mielke, Berry, and Johnson in 1976 [971]; the Fisher exact probability test by Soms and the Baker–Hubert test of ordering theory in 1977 [53, 1296]; the Agresti–Wackerly–Boyett test for $r \times c$ contingency tables in 1979 [8]; and a variety of permutation-generating algorithms by Page [1085], Boothroyd [180, 181], Bratley [206], Ord-Smith [1065], Phillips [1124], and Langdon [799] in 1967; Ord-Smith in 1968 [1067]; Chase in 1970 [247, 248]; Liu and Tang in 1973 [837]; Dershowitz in 1975 [344]; Rohl and Ives in 1976 [675, 1183]; Rohl in 1978 [1184]; and Payne and Ives in 1979 [1091].

1960–1979 4

Permutation methods were still not completely accepted in the early 1960s, even by some prominent and influential statisticians. The idea that permutation statistical tests constituted a standard against which conventional normal-theory tests could be evaluated continued to be questioned, and permutation tests were not regarded by many as legitimate alternatives to normal-theory tests. Recall that Frank Yates joined the Rothamsted Experimental Station in 1931, succeeding R.A. Fisher as the head of the Statistical Laboratory when Fisher left Rothamsted in 1933 to assume the post of Galton Professor of Eugenics at University College, London. R.A. Fisher passed away in Adelaide, Australia, in 1962 and in 1963 in a memorial issue of *Biometrics* commemorating the contributions of Fisher, Yates wrote that Fisher did not regard the regular use of permutation tests as reasonable, remarking "unfortunately tests of this nature, under the name of 'non-parametric tests', later came to have a certain vogue, which is not yet ended" [1474, p. 318]. Later, Yates reaffirmed his position, arguing that Fisher did not regard the regular use of randomization and other non-parametric tests as reasonable [1475, p. 782], citing Fisher from the last section of Chap. III in *The Design of Experiments* as saying:

> [permutation tests] were in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors. The utility of such non-parametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality [451, p. 48].

In defense of normal-theory tests, in 1964 Yates pointed out that in very small samples, the level of significance provided by a permutation test often will not agree with the level of significance provided by the corresponding normal-theory test, even on many samples of values from a normally-distributed population, and argued that "disagreement between the two tests . . . is not in itself evidence that the normal-theory test is inappropriate" [1474, p. 318]. An alternative point of view was provided in 1937 when Bernard Welch (q.v. page 74) published a paper on the use of Fisher's variance-ratio z test in randomized block and Latin square designs in which he compared permutation and normal-theory procedures, concluding

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9_4,
© Springer International Publishing Switzerland 2014

that the permutation procedure should be followed whenever the permutation and normal-theory tests yielded different results [1428]. In 1963 Yates addressed this recommendation by Welch, commenting: "[f]ortunately practical experimenters have never taken this suggestion seriously" [1474, p. 318].[1]

In spite of the resistance to permutation statistical methods, work continued in the field and much progress was recorded between 1960 and 1979. Interestingly, an area that contributed significantly to the growth of permutation tests during this era was not statistical, but rather occurred in an area only indirectly related to statistics at that time: computer science.

4.1 Overview of This Chapter

As is readily apparent, permutation statistical methods are computationally-intensive and ultimately depend on the efficient generation of permutation sequences. In the case of exact permutation tests, all possible permutation sequences are generated, but for Monte Carlo (resampling) permutation tests only a random sample of permutation sequences is required. Although the first explicit description of computer algorithms for the generation of permutation sequences was given by Tompkins in 1956 [1364], many algorithms were presented for the generation of permutation sequences in the period from 1960 to 1979, each touting increased speed, efficiency, or both.

Early in this period in 1961, C.R. Rao published a non-computer procedure for the generation of pseudorandom permutation sequences using a table of uniform pseudorandom numbers [1154]. Following the publication by Rao, many computer-based algorithms for permutation sequences were developed. Among them were sequence algorithms published by Coveyou and Sullivan [290], Wells [1435], Howell [658], Trotter [1372], Peck and Schrack [1112], Johnson [693], Heap [608], Durstenfeld [367], Sag [1213], Boothroyd [178, 180, 181], Bratley [206], Langdon [799], Robinson [1177], Ord-Smith [1065, 1067], Chase [247, 248], Dershowitz [344], Fike [432], Ives [675], Woodall [1469], Rohl [1184], and Payne and Ives [1091]. In addition, Ord-Smith in 1970 and 1971 [1068, 1069], Rabinowitz and Berenson in 1974 [1149], Sedgewick in 1977 [1242], and Lipski in 1979 [832] provided extensive summaries of the literature on the generation of permutation sequences in this period.

While computer algorithms to generate permutation sequences were important, other researchers turned their attention to computing exact probability values for established statistical tests. Gregory [553] and Tritchler and Pedrini [1371], for example, confined their applications to the Fisher exact probability test for 2 × 2 contingency tables, while Agresti and Wackerly [7], Agresti, Wackerly, and Boyett [8], Fleishman [466], Howell and Gordon [657], and March [890] attempted to

[1] Authors' note: it is abundantly evident from reading the many publications of Frank Yates that although he contributed significantly to the literature of permutation methods, he considered normal-theory tests as sacrosanct.

extend the Fisher hypergeometric procedure to contingency tables that were larger than 2×2, and other researchers applied permutation procedures to, for example, the Pitman test for two independent samples [30], the F test for completely randomized designs [52], the F test for randomized block designs [268], the chi-squared test for goodness of fit [1150], the Kruskal–Wallis analysis of variance rank test [779], and alternative choices of rank scores [932, 944]. On the topic of choices of rank scores, in 1972 Mielke investigated the asymptotic behavior of two-sample linear tests associated with infinite classes of distinct rank-order statistic functions [932, 933, 987]. The study was motivated by the asymptotic behavior and tied-value moment adjustments for linear tests based on specific sums of distinct, squared, rank-order statistic functions; see also papers by Taha in 1964 [1339], Mielke [931] and Grant and Mielke in 1967 [545], and Duran and Mielke in 1968 [366]. Finally, in 1969 Edgington provided permutation procedures and examples for an extensive inventory of statistical tests [391, pp. 93–159], and 10 years later Boyett published an important resampling algorithm for $r \times c$ contingency tables [199].

In 1976 Mielke, Berry, and Johnson [971] introduced multi-response permutation procedures (MRPP), techniques designed especially for data-dependent permutation methods per se, in contrast to permutation alternatives to standard statistical tests. Based on ordinary Euclidean distances rather than the squared Euclidean distances of conventional tests, MRPP provided highly robust, distribution-free, Euclidean-distance-based permutation alternatives for analyzing classical experimental designs that normally employed such established tests as analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA) [940, 978].

During the period from 1960 to 1979, researchers were focused on defining efficient methods for calculating probability values using existing computing machinery. Computing inefficiencies were largely due to inadequate numerical algorithms, low computer clock speeds, small and slow core memories, and inefficient data transfers. Mielke, Berry, and Johnson [971] and Mielke [936] pioneered moment-approximation permutation procedures implemented with the use of symmetric means, introduced by Tukey in 1950 [1375], and provided the exact first three moments of a continuous distribution that approximated the underlying discrete permutation distribution. Since asymptotic invariance procedures did not exist for many cases of MRPP, the three-moment approximation was essential for most cases when no asymptotic invariance procedure, such as normality, existed [220]. The moment-approximation permutation procedure immediately eliminated many of the computing difficulties that had plagued the computation of permutation probability values, provided an approximation to the underlying permutation distribution, and circumvented the extensive calculations of an exact permutation approach.

4.2 Development of Computing

The invention and development of the modern computer is one of the seminal events of humankind, ranking alongside the inventions of movable type and mechanical timepieces in advancing civilization. The ability to compute and to search for information accurately and efficiently was a major driver in transforming the developed world from an inefficient, error-plagued, uninformed society to one that was efficient, knowledgeable, and technologically sophisticated.

In the early years, computers allowed computations to be done faster and more precisely. Thus, tasks that formerly required several hours took only minutes on an early computer, such as ENIAC, EDSAC, MANIAC, or the Harvard Mark I—a *quantitative* difference in that complex problems could be solved faster, efficiently, and more accurately. Later, computers could compute in a few seconds what would formerly have taken 100 people 100 years to calculate. Thus, problems could be solved that could previously be only imagined—a *qualitative* difference in that problems that were impossible to solve could now be worked out in a few minutes. As Kenneth Appel was famously quoted as saying, "[w]ithout computers, we would be stuck only proving theorems that have short proofs" [1074, p. A19].[2]

Miniaturization of computer components and the development of the desktop computer made the computer more portable and more accessible to the average citizen. Along with accessibility and convenience, miniaturization led to greater precision. Today, high-speed computing is well within an individual's grasp with desktops, laptops, tablets, notebooks, netbooks, pads, and pods widely available at a reasonable cost.

Given the computationally-intensive nature of permutation methods, it took the development of high-speed computers for permutation-based statistical tests to achieve their potential. Thus, the parallel development of permutation tests and computing is an essential part of the chronology of permutation methods [695]. While this is not the proper place for a history of computing, some notable highlights of the development of computing between 1960 and 1979 are important for understanding the advancement of permutation statistical methods, especially those related to computing speed.

As Thisted and Velleman noted in 1992, statistical practice has long combined mathematical theory, methodological research, and applications to scientific problems [1353, p. 41]. Over time, as computers became more powerful and more accessible to researchers, they came to play an increasingly important role in all three areas. Further advances in computational power motivated the development of new statistical methods, such as permutation methods. Thisted and Velleman expressed it very succinctly when they wrote in 1992:

> [c]omputational advances have changed the face of statistical practice by transforming what we do and by challenging how we think about scientific problems [1353, p. 41].

In the 1940s, computing was called "automatic computation" and in the 1950s, "information processing." In the 1960s, computing acquired the name "computer science" in the United States and "informatics" in the United Kingdom. By the 1980s, computing was comprised of a complex of related fields, including computer

[2] Kenneth Appel and Wolfgang Haken used an IBM 370 mainframe at the University of Illinois to solve the four-color map problem in 1977.

science, informatics, computer engineering, software engineering, numerical analysis, and information technology. As noted by P.J. Denning, by 1990 the term "computing" had become the standard for referring to this core group of disciplines [343].

The leaders of the field struggled with the essential identity of computing from the very beginning. In the 1960s it was argued that computing was unique among all the sciences in its study of information processes. In the early 1970s, computing came to stand for algorithmic analysis and the catch phrase at this time was "computer science equals programming." In the late 1970s, computing was redefined as the automation of information processes. Finally, in the 1980s, the view was adopted that computing was not only a tool for science, but a new method of thought and discovery in science [343]. For the most part, permutation statistical methods developed by subscribing to this latter view.

Prior to 1960 computers were large, slow, expensive, and in large part their use was restricted to military and industrial applications. For example, consider the SAGE (Semi-Automatic Ground Environment) computer system that was initiated in the late 1950s, became operational in 1963, and served until 1983. The SAGE system used 30 large mainframe Whirlwind II computers built by IBM to coordinate the United States air defense systems [370, pp. 310,330]. Each Whirlwind II computer was 50 ft wide and 150 ft long, weighed 250 tons, and contained 60,000 vacuum tubes.[3] The SAGE system was the largest, heaviest, and most expensive computer system ever built, yet the computing power of each Whirlwind II computer was less than that of a single netbook computer of today [1243].

No account detailing the development of computing in this period would be complete without mention of Bill Joy, the co-founder of Sun Microsystems, the author of the vi editor, and the developer of csh, the C shell for UNIX platforms.

Bill Joy

William Nelson (Bill) Joy Jr. graduated from high school at age 15 in Farmington Hills, Michigan, and entered the University of Michigan, graduating with a B.S. degree in computer science in 1975. As late as the 1960s, computers were the size of small rooms and were quite rare. What is more, even if you could find a computer it was difficult to gain access to it, and if you could gain access, renting time on it could cost several thousands of dollars an hour [515, 702, 1394].

Programming, at the time, meant working with cardboard punch cards, with many programs consisting of hundreds, sometimes thousands, of cards. Since mainframe computers could handle only one task at a time

(continued)

[3] The Audion vacuum tube was invented by electrical engineer Lee de Forest and patented on 25 October 1906.

(batch-processing), the computer operator scheduled an "appointment" for a specified job and it might take hours, or even a day, to get a job run and returned. The University of Michigan, however, was one of the first universities in the world to switch from batch-processing to time-sharing. At the time that Joy was there, the University of Michigan possessed sufficient computing power that a hundred people could be logged on to the university mainframe and programming simultaneously. The University of Michigan Computer Center was in the North Campus where Joy lived. Joy had 24/7 access to the Computer Center and, through a bug in the Computer Center software, Joy was able to exploit the system and program without incurring any computing charges [515].

In 1975 Joy entered the University of California at Berkeley, graduating in 1979 with an M.S. degree in electrical engineering and computer science. While at Berkeley, Joy updated the department's UNIX operating system and won a contract to adapt the Berkeley version of UNIX for a project called "the Internet" from the United States Department of Defense Advanced Research Projects Agency (DARPA). Joy's development group adapted and reinvented two networking protocols: TCP (Transmission Control Protocol) and IP (Internet Protocol), and in 1976 Joy developed the vi editor for UNIX platforms. In 1982 Joy joined Vinod Khosla, a graduate of Stanford University, Scott McNealy, and Andreas Bechtolsheim to found Sun Microsystems, Incorporated, and to develop SUN (Stanford University Network) workstations. In 1991 Bill Joy relocated to Aspen, Colorado, where presently he works on assorted projects for Sun Microsystems under the rubric Aspen Smallworks, located high above Aspen in the shadows of Smuggler, Bell, and Shadow Mountains [1021, pp. 325–326]. Joy retired from Sun Microsystems as vice president of research and development on 9 September 2003.

As with Bill Joy, no account of the development of computing in this period would be complete without mention of Bill Gates, who with his long-time friend Paul Allen co-founded Microsoft, Incorporated, who also co-founded with his wife Melinda French Gates the Bill & Melinda Gates Foundation to reduce inequities in the United States and around the world, and who is the author of *The Road Ahead* first published in 1995 [497] and *Business @ the Speed of Thought* first published in 1999 [498].

Bill Gates

William Henry (Bill) Gates III was born on 28 October 1955 in Seattle, Washington. When Gates was 13, his parents removed him from Seattle's public schools and enrolled him in the seventh grade at Seattle's Lakeside

(continued)

School, an exclusive, all-boys, preparatory, private school that catered to Seattle's elite families located on 30 acres near the Jackson Park Golf Course in north Seattle [515].[4]

In 1968, The Mother's Club at Lakeside School raised and donated 3,000 dollars to purchase a computer terminal and computing time for the school; actually, an Teletype ASR-33 (for Automatic Send and Receive), which was a time-sharing terminal with a direct link to a computer in downtown Seattle; actually, a GE-634 mainframe built by General Electric. It is striking that Bill Joy got the early opportunity to learn programming on a time-share system as a freshman at the University of Michigan in 1971, while Bill Gates learned programming on a time-share system as an eighth-grade student at Lakeside School in Seattle in 1968 [496].

Gates spent countless hours programming at Lakeside, then was able, through the mother of another student at Lakeside, to acquire free computer time at the Computer Center Corporation (C-Cubed) on weekends, where a DEC PDP-10 resided. After C-Cubed went bankrupt, Gates found free computer time at Information Sciences, Incorporated, in exchange for working on software to automate company payrolls. As Gates once remarked, he had better exposure to software development than anyone else at that time. In the fall of 1973, Gates enrolled at Harvard University, having scored 1,590 out of a possible 1,600 on the College SAT test. It was at Harvard that Gates shared a dormitory room with Steven Anthony (Steve) Ballmer, who succeeded Gates as CEO of Microsoft in January of 2000. (On 23 August 2013, Ballmer announced his pending retirement as CEO of Microsoft.) In 1975, Gates dropped out of Harvard to found Micro-Soft (the hyphen was dropped after 1 year) with Paul Gardner Allen, a long-time childhood friend from Lakeside School who had dropped out of Washington State University in 1974 and moved to Boston to work for Honeywell as a computer programmer. After a somewhat shaky beginning, Microsoft's growth exploded between 1978 and 1981. In 1981 Gates and Allen incorporated Microsoft with Gates as president and chairman of the board and Allen as executive vice-president. In 1986, Gates took Microsoft public and in 2008 Gates transitioned out of a day-to-day role in Microsoft to spend more time at the Bill & Melinda Gates Foundation that was founded in 2000 [496, 515].

[4] The class of 1971 was Lakeside's last as an all-boys school; it merged with St. Nicholas, an all-girls school, to be co-educational that fall [13, p. 51].

Paul Allen

Paul Gardner Allen was born on 21 January 1953 in Seattle, Washington. Like Bill Gates, Allen attended Lakeside School, although he was two grades ahead of Gates. And like Gates, it was on the ASR-33 that Allen leaned to program. After graduation, Allen went to Washington State University in Pullman, Washington, where he was a member of Phi Kappa Theta (ΦΚΘ) fraternity. In 1974 Allen left Washington State three semesters shy of graduating to join Honeywell Corporation in Boston, Massachusetts, as a programmer. This put him close to Gates, whom he persuaded to leave Harvard and found Microsoft.

Paul Allen was diagnosed with Hodgkin's lymphoma in 1982. Although the cancer was successfully treated, he did not return to Microsoft and in 2000 he resigned from the Microsoft Board of Directors. Presently Paul Allen, in addition to his many business interests, is the owner of the Seattle Seahawks of the National Football League (NFL) and the Portland Trailblazers of the National Basketball Association (NBA). In 2011 Allen published *Idea Man: A Memoir by the Cofounder of Microsoft* [13].

In the late 1960s and early 1970s, mainframe computers became widely available to researchers at major research universities. In 1962 the LINC (Laboratory INstrument Computer) began processing data in the Lincoln Laboratory at the Massachusetts Institute of Technology to assist with biomedical research. The LINC was a small, stored-program, digital, 12-bit, 2,048-word computer designed to accept analog as well as digital inputs directly from experiments [1350]. In 1963 Douglas Engelbart invented the mouse in his research lab at the Stanford Research Institute SRI and a patent was issued in 1967. In September of 1964 the Control Data Corporation (CDC) introduced the first supercomputer, the CDC 6600, designed by Seymour Roger Cray and James Edward Thornton in Chippewa Falls, Wisconsin.

In 1960 Kenneth Iverson and Adin Falkoff at IBM created APL (A Programming Language) based on a non-conventional notational scheme that Iverson had created in 1957 while a faculty member at Harvard University. APL is an interpretive language based on a unique non-standard character set composed of symbols rather than words,[5] and has only one recursive precedence rule: all operators have equal precedence and all operators associate right to left.[6,7] The first personal computer implementation of APL was on the Intel 8008-based MCM/70 (Micro Computer Machines/70) personal computer in 1973.

In 1963 John George Kemeny (originally, Kemény János György) and Thomas Eugene Kurtz, both in the Mathematics Department of Dartmouth College in

[5] Some representative APL symbols are: ⍞, ⍗, ⍝, ∆, ⊛, and ⌹.

[6] For examples of APL statistical programs, see [109, 110, 124, 128, 132].

[7] A unique feature of APL is that any value divided by itself is equal to one, including zero divided by zero.

Hanover, New Hampshire, developed the BASIC (Beginners All-purpose Symbolic Instruction Code) computer language, which gave Dartmouth undergraduate students easy access to computing.[8] In 1965 the PDP-8, made by Digital Equipment Corporation (DEC), made its début and became the first microcomputer success; price: $18,000. In 1966 Maurice George Kendall, commenting on electronic computers, concluded that "for most practical purposes the out-of-core memory storage . . . is unlimited" and projecting ahead said "the process and access times of the next generation of computers will be reckoned in nano-seconds" [737, p. 1].

In 1969 the Department of Defense (DOD) established the first computer network, ARPAnet, and in 1971 ARPAnet transmitted the first email message.[9] In the period between 1969 and 1973, Dennis MacAlistair Ritchie and Kenneth Lane Thompson developed the UNIX operating system at Bell Laboratories (now, Alcatel–Lucent) in Murray Hill, New Jersey. Originally written in assembly language, the UNIX operating system was rewritten in C, a new general-purpose computer programming language developed by Ritchie and Thompson for use with the UNIX operating system. Subsequently, C was introduced to the public in 1978 with the publication of *The C Programming Language* by Brian W. Kernighan and Dennis M. Ritchie. In 1975 *Popular Electronics* put the Altair 8800 computer kit on its January cover and its maker, Micro Instrumentation Telemetry Systems (MITS), was flooded with requests; memory was only 256 bytes.

Somewhat prior to the introduction of the C programming language in 1978, a so-called "canned" Statistical Package for the Social Sciences (SPSS) was released in 1968 by Norman H. Nie, Hadlai (Tex) Hull, and Dale H. Bent. Development of SPSS began at Stanford University where Nie was a doctoral candidate in political science, Bent was a doctoral candidate in operations research, and Hull was a recent graduate of Stanford University with an MBA degree. SPSS incorporated in 1975, establishing its headquarters in Chicago, Illinois, and was publicly traded in August of 1993 [1422]. In 1976 another statistical package called Statistical Analysis System, or SAS, was released by the SAS Institute. SAS had its birth as a statistical analysis system in the late 1960s. SAS grew out of a project in the Department of Experimental Statistics at North Carolina State University at Raleigh. This project led to the formation of the SAS Institute in 1976 [1422]. SAS was originally developed by Anthony J. Barr, James H. Goodnight, John P. Sall, and Jane T. Helwig, in addition to a number of other early contributors. In the same year, 1976, the S programming language was developed at Bell Laboratories under the direction of John Chambers and Trevor Hastie, along with Richard A. Becker, Alan Wilks, and William S. Cleveland. S was written in C as a higher-level programming

[8] For a history of the development of the BASIC computing language, see the 1978 recollections of Thomas Kurtz in the special issue of *ACM SIGPLAN Notices* on the history of programming [782].

[9] A precursor to the Internet, ARPAnet was the first operational packet-switching network and was created for the United States Defense Advanced Research Projects Agency (DARPA) in 1969. ARPAnet was decommissioned in 1990 when it was superseded by the National Science Foundation Network (NSFNET).

language with separate algorithms developed for different statistical procedures [618].[10]

As with Bill Joy, Bill Gates, and Paul Allen, no history of computing in this period would be complete without mention of Gordon Moore, co-founder of Intel Corporation with Robert Noyce and Andrew Grove, and author of Moore's law [1002].

G.E. Moore

Gordon Earl Moore received his B.S. degree in chemistry from the University of California at Berkeley in 1950 and his Ph.D. in chemistry and physics from the California Institute of Technology in 1953. In 1957 he co-founded Fairchild Semiconductor with Julius Blank, Victor Grinich, Jean Hoerni, Eugene Kleiner, Jay Last, Robert Noyce, and Sheldon Roberts, known as the "traitorous eight" because they left William Bradford Shockley and the Shockley Semiconductor Laboratory to form their own company, Fairchild Semiconductor, in 1957 [103, Chap. 5]. In 1965 Moore published a short article in the 19 April issue of *Electronics* with the title "Cramming more components onto integrated circuits" [1002]. In this 1965 article Moore described a trend in the history of computing where the number of transistors that could be placed on an integrated circuit had doubled every year. He initially projected that the doubling would continue every year, but later revised the projection to doubling every 2 years [13, p. 2]. According to Moore, the trend later was labeled "Moore's Law" by computer scientist Carver Mead at the California Institute of Technology. The trend has been maintained more or less consistently for over 50 years. In July 1968 Moore left Fairchild Semiconductor and founded Intel Corporation with partners Robert Noyce and Andrew Grove [103, Chap. 7]. Moore retired from Intel in 1997.

As Michael Kanellos has related, Moore once extrapolated that if the car industry followed the same rules of progress, cars would get 100,000 miles per gallon, travel at millions of miles per hour, and be so cheap that it would cost less to buy a Rolls–Royce than to park it downtown for a day. However, as a friend pointed out, Moore also said, "[the car] would only be a half-inch long and a quarter-inch high" [708] (Moore, quoted in Seel [1243, p. 15]).

Beginning in 1975 with the success of Paul Allen, who at the time was a Honeywell programmer in Boston, Massachusetts, and Bill Gates, a freshman at Harvard University, who together wrote an interpreter for a subset of BASIC

[10] S-PLUS® is a commercial implementation of the computing language S and was first produced in 1988 by Statistical Sciences, Incorporated, a Seattle-based start-up company founded by R. Douglas Martin, a professor of statistics at the University of Washington, Seattle.

commands to the Altair 8800 personal computer, BASIC accelerated the personal computer revolution [598]. By the end of this period in 1979, personal computers, although not common, were available to many researchers, with the PDP-8, the first affordable mini-computer, introduced by the Digital Equipment Corporation in 1963; the Altair computer, the very first full-flexed personal computer on the market, introduced in 1975; and the Commodore PET introduced in 1977. In 1976 Steven Paul Jobs and Stephen Gary Wozniak, two college dropouts,[11] released the Apple I computer, which they had developed in the garage belonging to Jobs' parents. A year later in 1977, Jobs and Wozniak introduced the Apple II computer that included color graphics and housed its electronics inside a plastic case. The Apple II soon became the first mass-marketed personal computer.

During this same period, the speed of computing increased greatly. For example, in 1971 Intel introduced the 4004 microprocessor with 2,300 transistors and a clock speed of 108 KHz, but by 1979 the Intel 8088 microprocessor with 29,000 transistors was running at a speed of 5 MHz. By the mid-1970s, John Kemeny, who was then president of Dartmouth College, was quoted as saying that "the average undergraduate at Dartmouth with a pocket calculator was holding more computing power in his left hand than existed in the entire world just 15 years earlier," and in comparing his experiences as a young mathematician working in the theoretical division of the Manhattan Project at the Los Alamos National Laboratory in 1946 with those of a Dartmouth undergraduate in 1975, was quoted as saying that "[i]t took twenty of us working 20 h a day for an entire year to accomplish what one student can now do in an afternoon." In retrospect, the speed of computing increased greatly between 1960 and 1979, paving the way for the rapid development of permutation statistical methods.

4.3 Permutation Algorithms and Programs

Exact permutation statistical methods ultimately depend on the generation of the $n!$ possible permutations of the n consecutive integers from 1 to n (q.v. page 4). Alternatively, resampling-approximation permutation methods depend on the Monte Carlo generation of a random subset of the $n!$ possible permutations of the n consecutive integers from 1 to n. In both cases, the permutation sequences are used as subscripts to the observed measurement values so that the values can be shuffled in all $n!$ possible ways for an exact permutation analysis, or so that the $n!$ possible ways can be randomly sampled in a predetermined number of ways for a resampling-approximation permutation analysis. The 1960s and 1970s witnessed a proliferation of algorithms and programs to generate permutation sequences, each designed to be faster, more efficient, or more elegant than previous algorithms.

[11] Steve Wozniak eventually returned to college and completed his B.S. degree in Electrical Engineering and Computer Science at the University of California at Berkeley in 1987.

1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
3	4	2	4	2	3	3	4	1	4	1	3	2	4	1	4	1	2	2	3	1	3	1	2
4	3	4	2	3	2	4	3	4	1	3	1	4	2	4	1	2	1	3	2	3	1	2	1

Fig. 4.1 Example permutation sequences for the first four consecutive integers $\{1, 2, 3, 4\}$, where the sequences are to be read vertically

Figure 4.1 illustrates lexicographical sequences of $4! = 24$ permutations based on an initial sequence of the first four consecutive integers $\{1, 2, 3, 4\}$, where the permutations sequences are listed vertically.

Most of the algorithms published in this time period resulted in an exhaustive list of the $n!$ possible permutations of the consecutive integers from 1 to n, but several were designed to generate a random subset of all $n!$ possible permutations. It should be noted that, for the most part, these algorithms were not specifically designed with permutation statistical methods in mind, appearing as they did primarily in computer science journals. There are simply too many algorithms to examine here in detail, or even to list completely; however, a number of them should be mentioned, and a few deserve a thorough description.

In 1938 R.A. Fisher and Frank Yates published a method for obtaining random permutations of the consecutive integers from 1 to n utilizing tables of random digits [463]. Unfortunately, the process described by Fisher and Yates was inefficient, rejecting on average 75 % of the random numbers generated. In 1961 C.R. Rao presented a more efficient method of generating random permutations of the integers 1 to n for any n from a table of random digits that did not waste any random number generated [1154]. In 1962 M. Sandelius described a randomization procedure that consisted of distributing a deck of cards into ten decks using random decimal digits and repeating this step with each deck consisting of three or more cards [1220]. The procedure by Sandelius was essentially a special case of the general procedure described by Rao in 1961 [1154].

While the approaches described by Fisher and Yates [463], Rao [1154], and Sandelius [1220] utilized tables of random digits, in 1961 Coveyou and Sullivan described a computer algorithm utilizing a computer-based pseudorandom number generator that produced all permutations of the integers from 0 to n [290]. Also in 1961 Wells [1435], following on the work of Tompkins [1364], presented a scheme to generate all $n!$ permutations of n marks whereby each step consisted of merely transposing two marks (q.v. page 218), a procedure that was considerably faster than the Tompkins–Paige method presented by Tompkins in 1956[1364].

The transposition algorithm of Wells is typical of the permutation algorithms of this time. First, let P_n represent a permutation sequence of length n and place an arrow above every number in P_n, e.g., $\overleftarrow{1}\,\overleftarrow{2}\,\overleftarrow{3}\,\overleftarrow{4}$. Any number in P_n is considered

Fig. 4.2 Implementation of the Wells permutation algorithm by adjacent transposition with $n = 4$

p_1	p_2	p_3	p_4	M
$\overleftarrow{1}$	$\overleftarrow{2}$	$\overleftarrow{3}$	$\overleftarrow{4}$	4
$\overleftarrow{1}$	$\overleftarrow{2}$	$\overleftarrow{4}$	$\overleftarrow{3}$	4
$\overleftarrow{1}$	$\overleftarrow{4}$	$\overleftarrow{2}$	$\overleftarrow{3}$	4
$\overleftarrow{4}$	$\overleftarrow{1}$	$\overleftarrow{2}$	$\overleftarrow{3}$	3
$\overrightarrow{4}$	$\overleftarrow{1}$	$\overleftarrow{3}$	$\overleftarrow{2}$	4
$\overleftarrow{1}$	$\overrightarrow{4}$	$\overleftarrow{3}$	$\overleftarrow{2}$	4
$\overleftarrow{1}$	$\overleftarrow{3}$	$\overrightarrow{4}$	$\overleftarrow{2}$	4
$\overleftarrow{1}$	$\overleftarrow{3}$	$\overleftarrow{2}$	$\overrightarrow{4}$	3
$\overleftarrow{3}$	$\overleftarrow{1}$	$\overleftarrow{2}$	$\overleftarrow{4}$	4
$\overleftarrow{3}$	$\overleftarrow{1}$	$\overleftarrow{4}$	$\overleftarrow{2}$	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\overleftarrow{2}$	$\overleftarrow{1}$	$\overrightarrow{4}$	$\overrightarrow{3}$	4
$\overleftarrow{2}$	$\overleftarrow{1}$	$\overrightarrow{3}$	$\overrightarrow{4}$	0

to be in an "active state" when the adjacent number in the arrow direction of the number is smaller than the number itself. Thus, the numbers $\overleftarrow{2}$, $\overleftarrow{3}$, and $\overleftarrow{4}$ are in an active state in this example [cf. 1492]. The Wells adjacent transposition permutation algorithm can be described in three simple steps [1492]:

Fig. 4.2 illustrates implementation of the Wells permutation algorithm by adjacent transposition with $P_4 = \{\overleftarrow{1}\ \overleftarrow{2}\ \overleftarrow{3}\ \overleftarrow{4}\}$.

1. If there is no number in P_n in an active state, stop; otherwise go to Step 2.
2. Find the maximum number in P_n in an active state and label it M. Transpose M and the adjacent number in the arrow direction of M and go to Step 3.
3. Change the arrow direction of all the numbers in P_n that are larger than M and go to Step 1.

Random Number Generators

Prior to the widespread availability of computers, random numbers were obtained from mechanical devices such as well-stirred urns, dice, roulette wheels, or other instruments of chance, and the results were recorded in tables [355]. Up to 1955, tables of uniform pseudorandom number digits were published by Tippett [1362], Fisher and Yates [463], Kendall and Babington Smith [740], Peatman and Shafer [1111], Hald [570], Royo and Ferrer [1201], and Steinhaus [1316], among others. The total number of random digits in these tables ranged from 1,600 to 250,000. Beginning in 1947, the RAND Corporation compiled a million random digits by electronic simulation of a roulette wheel attached to a computer [225]. The device had 32 slots, of which

(continued)

12 were ignored; the others were numbered from 0 to 9 twice [1417].[12] The results were published in 1955 as *A Million Random Digits with 100,000 Normal Deviates* [1152].

Pseudorandom number generation by computer was in its infancy at this time when scientists began to explore efficient methods of obtaining a sequence of independent uniform random numbers with computer programs by deterministic functions with a specified distribution. Unfortunately, the sequences of digits produced at this time were not always very random; see articles by Behrenz in 1962 [92], Marsaglia and Bray in 1968 [895], and Grosenbaugh in 1969 [559]. A notable example was the "middle-square method" developed by John von Neumann at the Los Alamos National Laboratory that had a very short period and other weaknesses [1402]. As Knuth noted:

> [t]he authors of many contributions to the science of random number generation were unaware that particular methods they were advocating would prove to be inadequate [763, p. 173] (Knuth, quoted in Dodge [355, p. 331]).

The middle-square method of von Neumann for the generation of pseudorandom numbers is quite elementary and can be described in just four steps:

1. Define a seed number of length n.
2. Square the seed number to obtain a $2n$-digit number, adding leading zeroes if necessary.
3. The next pseudorandom number is the middle n digits.
4. Repeat as necessary to obtain the required number of pseudorandom numbers.

For the interested reader, Sowey [1298] provides an extensive bibliography on random number generation in the period 1927 to 1971; Niederreiter [1037], Rubenstein [1204], Ripley [1172], L'Ecuyer [807], and Tezuka [1348] provide surveys on uniform pseudorandom number generators; Teichroew provides a history of distribution sampling prior to the era of the computer [1344], and Knuth [763] provides a complete chapter of 177 pages on the generation of uniform pseudorandom numbers [355].

The year 1962 marked the beginning of a proliferation of computer-based permutation sequence generators. In 1962 Peck and Schrack presented algorithm PERMUTE [1112], which inspired algorithm PERM by Trotter [1372]. Although it was only 1962, Trotter noted that the excuse for adding PERM to the "growing pile of permutation generators" was that PERM offered an advantage in speed over previous algorithms [1372, p. 435]. Rather tongue-in-cheek, Trotter also noted that

[12] European roulette wheels have 37 slots (0–36), while American roulette wheels have 38 slots (0, 00, 1–36).

PERM "also has the (probably useless) property that the permutations it generates are alternatively odd and even" [1372, p. 435].

The Peck and Schrack [1112] and Trotter [1372] algorithms were quickly followed by a plethora of other algorithms to generate permutations or combinations, including algorithm PERMUTATION by Howell [658], PERMULEX by Schrack and Shimrat that produced permutations in lexicographical order [1238], PERMUTE by Eaves [371], COMBINATION by Kurtzberg [783], and an unnamed algorithm by Lotto [843], nearly all published in *Communications of the ACM* in 1962. In addition, in 1962 Shen [1258] published a new method to generate permutations and combinations in lexicographical order that proved superior to a well-known method of generation by addition utilized by Howell in algorithm PERMUTATION [659].

In 1963 Wolfson and Wright [1467], Wright and Wolfson [1470], and Mifsud [992, 993] presented algorithms to generate all possible combinations of n objects, Shen published algorithm PERLE that generated all possible permutations in lexicographical order [1259], Johnson published a paper on the generation of permutations by adjacent transposition wherein each permutation was derived from its predecessor by a single interchange of two marks in adjacent positions [693], and Heap presented methods for obtaining all possible permutations of a number of objects, in which each permutation differed from its predecessor only by the interchange of two of the objects [608]. The Heap algorithm was later described by Lipski in 1979 as "probably the most efficient method known" [832, p. 358].

The year 1964 turned out to be an important year for permutation sequence generators, in general, and random permutation sequence generators, in particular. First, Sag introduced an algorithm to generate all permutations of a set with repetitions [1213]. Second, Durstenfeld put forth procedure SHUFFLE that generated random permutations of a sequence $\{1, 2, \ldots, n\}$. The procedure by Durstenfeld was based on the shuffling method first described by Fisher and Yates in *Statistical Tables for Biological, Agricultural and Medical Research* in 1938, but more importantly it was popularized by Donald Knuth, Professor of Computer Science, Stanford University, when he included it in Volume 2 of his exhaustive four volume work on *The Art of Computer Programming* in 1969 [762].

D.E. Knuth

Donald Ervin Knuth is Emeritus Professor at Stanford University and author of *The Art of Computer Programming* (TAOCP), which consists of four volumes on *Fundamental Algorithms*, *Seminumerical Algorithms*, *Sorting and Searching*, and *Combinatorial Algorithms*. A fifth volume on *Syntactic Algorithms* is in preparation and expected in 2020. In 1999, *American Scientist* named *The Art of Computer Programming* as among the best twelve

(continued)

Table 4.1 Illustration of the Fisher–Yates and Durstenfeld shuffling procedures with $N = 6$

			Fisher–Yates		Durstenfeld	
LIMIT	I	J	ARRAY	STORE	SWAP	ARRAY
6	0.54	4	123456	4	4 ↔ 6	123456
5	0.46	3	123 56	34	3 ↔ 5	123654
4	0.82	4	12 56	634	4 ↔ 4	125634
3	0.37	2	12 5	2634	2 ↔ 3	125634
2	0.16	1	1 5	12634	1 ↔ 2	152634
1			5	512634		512634

physical-science monographs of the century [1007]. For completeness, the other eleven books were:

1. Paul Dirac, *Quantum Mechanics* (1930)
2. Albert Einstein, *The Collected Papers of Albert Einstein: The Swiss Years: Writings, 1902–1909* (1930)
3. Benoit B. Mandelbrot, *Fractals* (1977)
4. Linus Pauling, *Nature of the Chemical Bond* (1939)
5. Bertrand Russell and Alfred North Whitehead, *Principia Mathematica, Volumes 1, 2, and 3* (1910–1913)
6. Cyril Smith, *Search for Structure* (1981)
7. John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (1944)
8. Norbert Weiner, *Cybernetics* (1948)
9. Richard B. Woodward and Roald Hoffmann, *Conservation of Orbital Symmetry* (1970)
10. Albert Einstein, *The Meaning of Relativity* (1922)
11. Richard Feynman, *QED* (1985)

Durstenfeld's algorithm differed from that proposed by Fisher–Yates in that instead of removing elements of ARRAY to storage array STORE, he swapped each selected element with the last unswapped element at each step. The algorithm by Durstenfeld is known as an *in situ* procedure as it shuffled the numbers of the array in place, rather than storing them elsewhere.[13] Table 4.1 illustrates the Fisher–Yates and Durstenfeld shuffling procedures for the sequence {123456}.

[13] Another difference between the Fisher–Yates and Durstenfeld shuffling procedures is that Fisher–Yates obtained their pseudorandom numbers from tables of random digits, while Durstenfeld used a computer-based pseudorandom number generator.

The Fisher–Yates Shuffle

The Fisher–Yates shuffle was designed to generate a random permutation of the consecutive integers $1, 2, \ldots, N$ and can be summarized in just a few steps:

1. Store the numbers from 1 to N in array ARRAY and set the value of LIMIT $= N$.
2. Choose a pseudorandom number, I, on $[0, 1)$ and scale it so it lies between 1 and LIMIT, which is N on the first selection. Denote the rescaled pseudorandom number J, where $J = \text{Int}(I \times \text{LIMIT}) + 1$.
3. Starting from the low end, remove the Jth remaining element of ARRAY and store it in array STORE.
4. Set LIMIT to LIMIT–1 and go to step 2, repeating until all the elements of ARRAY numbers have been moved to array STORE.

In 1965 Hill and Pike designed an algorithm for computing tail-area probability values for 2 × 2 contingency tables that was based on the exact method for fixed marginal frequency totals [622]. It was interesting because it provided a one-tailed probability value by summing the individual probability values equal to or less than the observed probability value, and then provided two quite different two-tailed exact probability values. One two-tailed probability value was obtained from the sum of the one-tailed probability value and a probability value calculated in similar fashion from the second tail. The second two-tailed probability value was obtained by including in the second tail all those terms that gave an inverse odds-ratio statistic as least as great as the odds-ratio statistic for the observed table.[14]

In 1967 algorithms to generate permutation sequences were presented by Page [1085], Boothroyd [178–181], Bratley [206], Ord-Smith [1065], Phillips [1124], and Langdon [799, 800]. The procedure by Langdon prompted a criticism by Ord-Smith [1066] and a defense by Rodden [1181], both based on Langdon's use of a rotational scheme designed to capitalize on the hardware design of computers of the time instead of the more conventional transpositional scheme. In 1991, Rohl showed that the pseudo-lexicographical algorithm of Ord-Smith [1065] was essentially equivalent to the Tompkins–Paige algorithm, given by Peck and Schrack [1112].[15]

In 1968 Ord-Smith introduced algorithm BESTLEX based on transpositions that produced all $n!$ permutations of n marks in lexicographical order [1067], and Plackett published an extensive article on permutations in which he described an algorithm that minimized the amount of randomization necessary to generate

[14] Almost 20 years later, in 1984, Hill explained that he and Pike could not agree on how to compute the two-tailed probability value. Pike argued for the odds-ratio method and Hill for the first method. In the end, as Hill noted "our algorithm included both and gave the user the choice" [620, p. 452].

[15] Schrack is variously misspelled in the literature as Schrank [1185] and Schrock [270].

a random permutation [1136]. In addition, Plackett described the probability distributions for the number of digits required by standard methods of generation from a sequence of random digits [1136].

In 1970 Chase published algorithm TWIDDLE designed to generate combinations of M out of N objects, which was based on an unpublished procedure discovered by Leo Lathroum in 1965 [248]; see also a remark by Chase in 1970 [249, p. 368]. Algorithm TWIDDLE was the combination equivalent to the Johnson–Trotter permutation generator [693, 1372]. Also in 1970, Ord-Smith published Part 1 (of two parts) of an overview on the generation of permutation sequences in which he detailed algorithms based on the Tompkins algorithm, nested cycle algorithms, the Wells, Johnson, and Trotter algorithms, lexicographic algorithms, and pseudo-lexicographic algorithms [1068]. Part 1 was quickly followed by Part 2 in 1971 [1069]. Here, Ord-Smith presented what he considered to be the six fastest general permutation algorithms; the article concluded with an extensive bibliography on published permutation sequence generators available at the time.

In 1971 Thomas presented an algorithm and FORTRAN subroutine for exact confidence limits for the odds-ratio statistic in 2 × 2 contingency tables [1355]. Assuming fixed marginal frequency totals, an iterative process was employed. In 1975 Thomas extended his work on the odds-ratio statistic to include exact and asymptotic methods for a series of 2 × 2 contingency tables [1356]. He provided an option for computing exact one- or two-tailed confidence limits for the odds-ratio statistic. As Thomas noted, since this was a discrete problem it was not possible to obtain a predetermined confidence interval of exactly $1 - \alpha$, but rather $1 - \alpha'$, where $\alpha' < \alpha$ and α' depended on the fixed marginal frequency totals. Thus, the results were "exact" in the sense that the confidence limits were at least $1 - \alpha$ and, consequently, always conservative [1356, p. 425, fn. 1].

In 1973 Liu and Tang developed subroutine NXCBN in FORTRAN that generated all combinations of m out of n objects [837]. Also in 1973, Ehrlich presented four new combinatorial algorithms [403]. The four algorithms had in common the important property that they used neither loops nor recursion; thus, the time needed for producing a new configuration was unaffected by the size of the configuration. The listing of these four algorithms was followed by a more lengthy discussion on loop-free algorithms for generating permutations, combinations, and other combinatorial configurations [403].

In 1975 Dershowitz described a simplified loop-free algorithm for generating all $n!$ permutations of a set of n elements (q.v. page 4). This was a simplification of Ehrlich's loop-free version of Johnson's and Trotter's algorithms [693, 1372]. Each permutation was generated by exchanging two adjacent elements of the preceding permutation. Also in 1975, Bebbington presented a simple method of drawing a random sample without replacement that was essentially a Fisher–Yates shuffle of elements [90]. Finally in 1975, Fike described a new method for generating permutation sequences [432]. Timing experiments indicated that the method proposed by Fike was competitive with the interchange methods of Wells [1435], Johnson [693], and Trotter [1372].

In 1976 Rohl presented programming improvements based on recursion procedures for Fike's algorithm for generating permutations that improved the performance of Fike's algorithm by a factor of two [1183]. Also in 1976 Hu and Tien noted that when all items were distinct the algorithms developed by Johnson [693] and Trotter [1372] generated all permutations by adjacent transposition, but the method did not provide a solution if not all items were distinct. To this end, Hu and Tien proposed an algorithm to generate all permutations when not all items were distinct [661]. The algorithm of Hu and Tien was an extension of the 1963 algorithm of Johnson [693] and was based on a series of binary sequences.

In 1976 Ives introduced four new algorithms for generating the $n!$ permutations of n marks [675] (q.v. page 4). Performance checks by Ives showed superiority of the new algorithms over Boothroyd's implementation of the algorithm by Wells and Ehrlich's implementation of the Johnson–Trotter algorithm.

In 1977 Woodall noted that Fike's algorithm had proved to be one of the fastest known, but he was able to develop a new algorithm that was even faster [1469]. Woodall's algorithm LEXPERM was a lexicographic procedure where recursion was eliminated, yielding an algorithm with a very fast procedure time. In the same year, Buckles and Lybanon presented a new algorithm COMB to generate a random set of combinations of n items taken p at a time and arranged in lexicographical order [228]. Finally in 1977, Sedgewick produced an extensive survey of permutation generation methods in which he surveyed the numerous methods that had been proposed for permutation generation by computer, described the various algorithms that had been developed over the years in considerable detail, and implemented them in a modern ALGOL-like language [1242]. In addition, as Sedgewick noted, the paper was intended not only as a survey of permutation methods, but also as a tutorial on how to compare a number of different algorithms for the same computing task [1242, p. 137].

In 1978 two articles were published that are worth mentioning. The first article by Rohl provided a simple, general algorithm to produce arrangements of n marks taken r at a time, where the marks need not be distinct [1184]. Various procedures based on the new algorithm were presented by Rohl, some producing arrangements in lexicographical order, some not. As Rohl noted, more important than the algorithm itself was the technique involved in its implementation—the use of a procedure that contained within itself a second procedure that was highly recursive. Thus, the algorithm effectively simulated a nest of r loops by means of a recursive procedure that called itself r times [1184, p. 305].[16]

The second article, by Roy, evaluated permutation algorithms with special attention to those published since Ord-Smith's review of algorithms in 1970 and 1971 [1199]. These included new algorithms by Fike [432] and Ives [675] and

[16] It should be noted that most of these procedures were written in computing languages that did not permit recursion, such as FORTRAN; therefore, it was necessary to generate all combinations by simulating nested loops. The problem was addressed by Jane Gentleman in 1975 with subroutine ALLNR, written in FORTRAN, that generated a complete set of all ${}_NC_R$ combinations of N things considered R at a time using simulated nested loops instead of recursion [507].

improvements to previously published algorithms by Ehrlich [403], Lenstra [820], and Rohl [1183]. Roy noted that there were a number of different permutation sequences that were widely used: one by Ives, another by Wells, those that were lexicographic, and those that were pseudo-lexicographic.[17] However, Roy found that all the different permutation sequences were generated from only two fundamental schemes.

The first scheme generated the $n!$ permutations of the n marks from knowledge of the $(n-1)!$ permutations of the first $(n-1)$ marks. Each of the $(n-1)$ permutations yielded n of the n-permutations [1199, p. 296]. The second scheme was described by Roy as follows. Suppose a procedure can generate only the $(k-1)!$ permutations of $k - 1$ marks $(k < n)$. The $k!$ permutations of the first k marks can then be generated by repeating the procedure k times by taking $(k - 1)$ of the k marks at a time and the remaining mark occupying the kth position.

Roy termed algorithms using the first scheme A-type algorithms, and those using the second scheme, B-type algorithms. Roy determined that the Ives procedure was the best A-type algorithm, the Wells procedure the best B-type algorithm, and that A-type algorithms were, in general, superior to B-type algorithms.

In 1979 Payne and Ives reconsidered the 1973 Liu–Tang combination enumeration algorithm that produced a cyclic sequence of combinations [1091]. While the Liu–Tang algorithm relied on generating combinations from marks, Payne and Ives considered pointers to the marks [1091].

Marks and Pointers to Marks

Combination and permutation sequences can be specified in two fundamental ways: marks and pointers to marks. First, the marks can be specified. For example, the combination $\binom{8}{5}$ can be written with 5 0s and $8 - 5 = 3$ 1s, e.g., {00000111}, which is the sum of binary 5 {00000101} and binary 3 {00000011}, yielding binary 8 {00001000}. Then, the original sequence of marks {00000111} can be systematically rearranged, e.g., {00001011}, {00001101}, {00001110}, and so on, always maintaining 5 0s and 3 1s.

Second, the serial location specified by pointers to the marks can be specified. For this example, 678, locating the positions of the 1s for {00000111}, i.e., the 1s are in positions 6, 7, and 8, from the left, and the 0s are in positions 1, 2, 3, 4, and 5, from the left. Then, the pointers for rearrangement {00001011} are 578, the pointers for rearrangement {00001101} are 568, and the pointers for rearrangement {00001110} are 567, and so on. Alternatively, the pointers can refer to the 0s: 12345, 12346, 12347, and 12348, respectively, but it is customary, and more efficient, to point to the less numerous marks.

(continued)

[17] In general, ordered permutations, such as lexicographic sequences, are of no consequence in permutation methods, either exact or resampling.

Fig. 4.3 Robertson's notation for an observed 2 × 2 contingency table

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d$

> Most combination generators use only the pointers to the marks rather than the marks themselves, as the pointers represent the ranks of one set of two combined ordered samples [1091].

Payne and Ives developed a pointer-programmed version of the Liu–Tang algorithm that greatly improved the speed of execution compared to the original coding, when k or $n - k$ was small as they appear in $\binom{n}{k}$. They compared their implementation not only to the Liu–Tang algorithm based on rearrangements of marks, but also with other combination enumeration generators by Kurtzberg [783], Chase [247], Bitner, Ehrlich, and Reingold [167], Mifsud [992], and Ehrlich [403, 404].

4.3.1 Permutation Methods and Contingency Tables

In work that was to prove to be a harbinger to the extensive contributions to come in the 1980s and 1990s, a number of articles were introduced on the computation of exact probability values for contingency tables and goodness-of-fit tests between 1960 and 1979. In 1960 Robertson published an article on programming Fisher's exact probability method of comparing two percentages [1174]. In this paper, Robertson described the application of a high-speed computer for determining the exact probability associated with the problem of comparing two percentages utilizing the Fisher–Yates exact probability method.[18] In programming the Fisher–Yates exact probability method, Robertson relied on stored logarithms of factorials. Robertson's notation for the cell frequencies and marginal frequency totals is given in Fig. 4.3 and the Fisher–Yates exact probability of any 2 × 2 contingency table was given by

$$P = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,(a+b+c+d)!}\ .$$

[18] The "high speed computer" in this case was a Royal McBee LGP-30. The Royal McBee Librascope General Purpose (LGP) computer was considered a desktop computer, even though it weighed 740 pounds, contained a 4,096 word magnetic drum memory and had a clock rate of 120 kHz.

Fig. 4.4 Feldman and Klinger's representation of an observed 2 × 2 contingency table

a_0	b_0	$a_0 + b_0$
c_0	d_0	$c_0 + d_0$
$a_0 + c_0$	$b_0 + d_0$	N

Robertson found that the computing time required for his program varied directly with the magnitudes of a and c, but was independent of the magnitudes of b and d. Consequently, the speed of the program depended largely on the number of division-multiplication cycles involved, which he calculated to be precisely $(a + 1)(3a + 2c)/2$.

In 1963 Feldman and Klinger published an efficient method for calculating the Fisher–Yates exact probability test for 2 × 2 contingency tables [424]. By that time, however, Finney had already published tables in 1948 for the Fisher–Yates exact test with marginal frequency totals up to 15 [434], Latscha had extended Finney's tables for marginal frequency totals up to 20 in 1953 [804], Armsen had extended Latscha's tables for marginal frequency totals up to 50 in 1955 [34], and in 1963 Finney, Latscha, Bennett, and Hsu published *Tables for Testing Significance in a 2 × 2 Contingency Table* [439]. Also, Finney's tables had already been incorporated into the widely-distributed *Biometrika Tables for Statisticians* by E.S. Pearson and H.O. Hartley in 1954 [1101]. However, Feldman and Klinger felt the need for a solution that fell outside the scope of the published tables [424, p. 289]. They argued that the tabled values suffered from two limitations. First, the tables reported critical values only for selected levels of significance, e.g., 0.05, 0.025, 0.01. Second, for $N > 30$ the tables listed critical values only for cases with equal marginal frequency totals [553, p. 698].

Given a 2 × 2 contingency table as illustrated in Fig. 4.4, the procedure suggested by Feldman and Klinger was to apply the usual formula for the hypergeometric probability value,

$$P_0 = \frac{(a_0 + b_0)!\,(a_0 + c_0)!\,(b_0 + d_0)!\,(c_0 + d_0)!}{N!\,a_0!\,b_0!\,c_0!\,d_0!} \;,$$

only to the observed table. Since

$$P_1 = \frac{a_0\,d_0}{b_1\,c_1} P_0$$

and, in general,

$$P_{i+1} = \frac{a_i\,d_i}{b_{i+1}\,c_{i+1}} P_i \;,$$

the solution proposed by Feldman and Klinger was a recursive procedure based on the observed probability value, where a researcher need only determine P_0 and then multiply it and each subsequent P_i by the product of the ith diagonal that was reduced, divided by the product of the $(i+1)$th diagonal that was increased. Then, summing the probability values that were as or more extreme than the observed probability value yielded the appropriate probability estimate for the observed contingency table [424, p. 291]. However, Johnson noted that while Feldman and Klinger recommended that the data be arranged such that $a_0 \leq b_0$, c_0, d_0, this was in error as it provided the complement of the desired probability value [689].

In 1964 Arnold investigated the multivariate generalization of Student's t test for two independent samples [35]. The first four permutation cumulants were determined for a statistic that was a simple function of Hotelling's T^2 test given by

$$t = \frac{T^2}{(m-1)+T^2} ,$$

where m was the number of blocks, and applied by Arnold to samples from bivariate normal, rectangular, and double exponential distributions. The samples examined ranged in size from $n = 48$ for $m = 4$ to $n = 800$ for $m = 8$. The results suggested that a test utilizing Hotelling's generalized T^2 statistic, when applied to non-normal data, was not likely to be biased by more than 1 or 2 percentage points at the 5 % level of significance.

In 1968 Hope introduced a simplified Monte Carlo test procedure for significance testing [649]. Noting that exact permutation tests were unnecessarily complicated due to the excessive number of permutations required, Hope advocated Monte Carlo (resampling) test procedures with smaller reference sets than required by exact permutation tests. Hope was able to demonstrate that the necessary number of Monte Carlo permutations could be determined from the level of significance adopted. For additional articles with a similar theme in this period, see a 1977 article by Besag and Diggle [164] and a 1979 article by Marriott [894].

In 1969 Kempthorne and Doerfler published a paper examining the behavior of selected tests of significance under experimental randomization [725]. They selected three tests for a matched-pairs design and concluded that the Fisher randomization test was to be preferred over the Wilcoxon matched-pairs rank-sum test, which in turn, was to be preferred over the sign test. All comparisons were based on Monte Carlo test procedures with 50 sets of randomly-generated data from eight distributions for experiments on 3–6 pairs of observations.

While the purported purpose of the paper was to compare matched-pairs designs, the paper actually contained a great deal more. First, Kempthorne and Doerfler objected to the use of specified cut-off points for the significance level α, and to classifying the conclusion as being simply significant or not significant, i.e., less than or greater than α. They argued that the use of such a dichotomy was inappropriate in the reporting of experimental data as it resulted in a loss of information [725, p. 239]. Second, they objected to the common practice of adding very small values such as 10^{-100} to measurements so as to avoid ties when converting to ranks. They referred to this practice as "fudging" the data. Third, they suggested that the term "significance level" of a test be eliminated from the statistical vocabulary; see also a 2012 article by Megan Higgs on this topic in *American Scientist* [616]. Finally, they dismissed the assumption of random samples in comparative populations and praised randomization tests for their ability to answer the question "What does this experiment, on its own, tell us?" [725, p. 235].[19] For a concise summary of the Kempthorne and Doerfler paper, see Kempthorne [720, pp. 22–25].

In 1971 Zelen considered the problem of analyzing data arranged into $k \geq 2$ contingency tables, each of size 2×2 [1487]. The principal result was the derivation of a statistical test for determining whether each of the k contingency tables has the same relative risk. Zelen noted that the test was based on a conditional reference set and regarded the solution as an extension of the Fisher–Irwin exact probability test for a single 2×2 contingency table [1487, p. 129].[20]

In 1954 Cochran [260] had investigated this problem with respect to testing whether the success probability for each of two treatments was the same for every contingency table, recommending the technique whenever the difference between the two populations on a logit or probit scale was nearly constant for each contingency table. Note that the constant logistic difference is equivalent to the relative risk being equal for all k tables.[21]

Mantel and Haenszel had previously proposed a method very similar to Cochran's, except for a modification dealing with the correction factor associated with a finite population [887]. Zelen investigated the more general problem when the difference between logits in each table was not necessarily constant [1487]. The exact and asymptotic distributions were derived by Zelen for both the null and non-null cases.

4.4 Ghent and the Fisher–Yates Exact Test

No account of the analysis of contingency tables would be complete without mention of the work of Arthur Ghent, who in 1972 extended the method of binomial coefficients first proposed by Patrick Leslie in 1955 [510].

[19] For Kempthorne and Doerfler, while randomization tests are based on permutations of the observations, they reserved the term "permutation tests" for the comparison of random samples from unspecified distributions and "randomization tests" for the comparison of the material actually used in an experiment.

[20] Recall that Fisher in 1935 [452], Yates in 1934 [1472], and Irwin in 1935 [674] independently developed the exact permutation analysis of a 2×2 contingency table with fixed marginal frequency totals (qq.v. pages 25, 37, and 48). Thus, references to either the Fisher–Yates or the Fisher–Irwin exact probability test are quite common.

[21] In general, when considering multiple 2×2 contingency tables the relative risk for each table must be in the same direction, e.g., measures of relative risk such as odds-ratios must all be greater (less) than 1 and approximately equal in magnitude.

Fig. 4.5 Example Ghent observed 2 × 2 contingency table

1	7	8
6	6	12
7	13	20

A.W. Ghent

Born in Canada, Arthur W. Ghent earned his B.Sc. and M.A. degrees in zoology from the University of Toronto in 1950 and 1954, respectively, and his Ph.D. in zoology at the University of Chicago in 1960. Ghent worked as a forest ecologist with the Canada Department of Agriculture while he was a student and upon his graduation, joined the faculty at the University of Oklahoma to begin an academic career as Assistant Professor of Quantitative Zoology. In 1964 Ghent moved to the University of Illinois where he was appointed Assistant Professor of zoology, achieving the ranks of Associate Professor and Professor in 1965 and 1970, respectively. In 1973 he accepted an appointment as Professor, School of Medical Sciences, at the University of Illinois. Arthur W. Ghent retired from the University of Illinois in 1997 and passed away on 27 April 2001 in Urbana, Illinois, at the age of 73 [509].

In 1972 Ghent examined the literature on the alignment and multiplication of appropriate binomial coefficients for computing the Fisher–Yates exact probability test for 2 × 2 contingency tables with fixed marginal frequency totals [510]. In an exceptionally clear and cogent presentation, Ghent reviewed the method of binomial coefficients first proposed by P.H. Leslie in 1955 [821] and independently discovered by Sakoda and Cohen in 1957 [1216].

The method of binomial coefficients, as described by Leslie [821], was a computational procedure involving, first, the selection of the appropriate series of binomial coefficients; second, their alignment at starting points in accord with the configuration of integers in the observed contingency table; and finally, the multiplication of adjacent coefficients that constitute the numerators of the exact hypergeometric probability values of all 2 × 2 contingency tables equal to or more extreme than the probability of the observed contingency table, given fixed marginal frequency totals [510, pp. 18–19].

An example will illustrate the binomial-product method as described by Ghent. Consider the example observed 2 × 2 contingency table in Fig. 4.5, where it is only necessary to examine the first row, as the second row is redundant, given the fixed marginal frequency totals.

For Cell $(1, 1)$ in Fig. 4.5, the cell frequencies can vary from a minimum of 0 to a maximum of 7, the first column marginal frequency total. The possible cell frequencies for Cell $(1, 1)$ are listed in the first column of Table 4.2. On the other hand, the cell frequencies in Cell $(1, 2)$ can vary only from a maximum of 8 to a minimum of 1, and not from the column marginal frequency total of 13 down to

Table 4.2 Illustration of the Ghent method of binomial coefficients to obtain Fisher–Yates exact probability values for a 2 × 2 contingency table

Cell		Binomial coefficients					
(1, 1)	(1, 2)	$(p+q)^7$		$(p+q)^{13}$		Product	Probability
0	8	1	×	1,287	=	1,287	0.0102
1	7	7	×	1,716	=	12,012	0.0954
2	6	21	×	1,716	=	36,036	0.2861
3	5	35	×	1,287	=	45,045	0.3576
4	4	35	×	715	=	25,025	0.1987
5	3	21	×	286	=	6,006	0.0477
6	2	7	×	78	=	546	0.0043
7	1	1	×	13	=	13	0.0001
Total						125,970	1.0000

a minimum of 0, as the frequencies for Cell (1, 2) are constrained by the marginal frequency total of 8 in the first row i.e., the two cells, (1, 1) and (1, 2), must sum to the row marginal frequency total of 8. The corresponding possible cell frequencies for Cell (1, 2) are listed in the second column of Table 4.2.

Because the frequencies for Cell (1, 1) can vary over the entire range of 0 to 7, the full complement of binomial coefficients of $(p+q)^7$ is listed in the third column of Table 4.2. However, the binomial coefficients of $(p+q)^{13}$, which are $\{1, 13, 78, 286, 715, 1{,}287, 1{,}716, 1{,}716, 1{,}287, 715, 286, 78, 13, 1\}$, are constrained by the range of possible cell frequencies for Cell (1, 2), i.e., from 8 to 1. Since 9, 10, 11, 12, and 13 are not possible cell frequencies for Cell (1, 2), eliminate the first five terms from the binomial coefficients for $(p+q)^{13}$, i.e., 1, 13, 78, 286, and 715, and since 0 is not a possible cell frequency for Cell (1, 2), eliminate the last term from the binomial coefficients for $(p+q)^{13}$, i.e., 1. The remaining binomial coefficients are listed in the fourth column of Table 4.2. The required binomial coefficients can easily be obtained from the first $n+1$ terms of the expanded binomial series,

$$1 + \frac{n}{1!} + \frac{n(n-1)}{2!} + \frac{n(n-1)(n-2)}{3!} + \cdots + \frac{n!}{n!} = \sum_{i=0}^{n} \binom{n}{i} = 2^n$$

or, for small samples, from Pascal's triangle (q.v. page 185). The two binomial series are then multiplied together and totaled, as illustrated in the fifth (Product) column of Table 4.2. Dividing each binomial product by the total yields the exact probability values, as listed in the last (Probability) column of Table 4.2.

The procedure is also described in some detail in Chap. 1940–1959 in the section on Patrick Leslie (q.v. page 184). Ghent extended the Leslie procedure to 2 × 3 and 2 × c contingency tables, requiring 3 and c series of binomial coefficients, respectively. Finally, he extended these results to 3 × 3 and $r \times c$ contingency tables using a two-step procedure that collapsed the larger contingency tables into smaller tables, then reassembled the results.

Since the 1930s, some controversy has existed over how correctly to compute a two-tailed probability value for the Fisher–Yates–Irwin exact probability test for 2×2 contingency tables. One approach is to sum the tail probability values equal to or less than the observed probability value for the tail in which the observed table fell, then simply double that probability value (doubling rule); see for example, an article by D.J. Finney in 1948 [434, p. 146]. The second approach is to sum the tail probability values in the "observed" tail, then add to that sum the sum of the probability values equal to or less than the observed probability value in the other tail (Irwin's rule) [674].

The difference can be illustrated with the listings in Table 4.2. The observed table containing cell values $(1, 7)$ with a binomial product of 12,012 is in the upper tail of the distribution of products. Thus, its one-tail exact probability value is $(1{,}287 + 12{,}012)/125{,}970 = 0.1056$, since 1,287 is less than 12,012. Doubling that probability value yields a two-tailed probability value of $2 \times 0.1056 = 0.2112$. On the other hand, the binomial products less than 12,012 in the lower tail are 13, 546, and 6,006. Then, the two-tailed probability value is $(1{,}287 + 12{,}012 + 13 + 546 + 6{,}006)/125{,}970 = 0.1577$.

Ghent was unequivocal on this matter, noting that "it is the sum of the equally, or more, extreme probabilities *separately calculated in both tails* that is logically continuous with the procedure by which Freeman and Halton (1951) obtain probabilities for 2×3 and larger contingency tables in their extension of the Fisher exact test principle" [510, p. 20].[22]

4.5 Programs for Contingency Table Analysis

In 1973 Gregory developed a FORTRAN computer program for the Fisher–Yates exact probability test that yielded a one-tailed exact probability value [553]. In this article Gregory made the controversial statement that since the Fisher–Yates statistic was inherently one-tailed, "the derived probability is simply doubled to test a two-tailed hypothesis" [553, p. 697]. This, of course, is certainly true if the two sets of marginal frequency totals are identical, resulting in a symmetric probability distribution; otherwise, it is a subject of some considerable debate. On this matter, see also articles by Cormack [279, 280], Haber [564], Healy [604], Jagger [678], Lloyd [838], Mantel [884, 885], Plackett [1139], and Yates [1476].

In 1975 Tritchler and Pedrini published a computer program for the Fisher–Yates exact probability test that yielded a one-tailed probability value and could evaluate samples up to size $n = 500$ [1371]. Also in 1975, Hays presented a FORTRAN procedure for the Fisher–Yates exact probability test [603]. The program relied on logarithms of factorials and produced the exact probability value associated with the observed 2×2 contingency table, the exact probability values associated with each

[22] Emphasis in the original.

of the more extreme possible tables, and the two-tailed probability of observing a result as or more divergent than that in the observed 2 × 2 contingency table.

In 1978 J. Berkson published a controversial article titled "In dispraise of the exact test" [102] and a second article questioning whether the marginal frequency totals of the 2 × 2 contingency table contain relevant information regarding the table proportions [101]. In these two articles, Berkson disagreed with Fisher's assertion that the marginal frequency totals of a 2 × 2 contingency table were "ancillary statistics" and therefore the observed marginal frequency totals provided no information regarding the configurations of the body of the table. However, Berkson's argument was incorrect. Berkson compared one-sided probability values from the Fisher–Yates exact probability test for 2 × 2 contingency tables with the normal test for the nominal significance levels 0.05 and 0.01, i.e., an exact versus asymptotic comparison. He showed that the effective level was closer to the nominal level with the normal test than with the exact test and concluded that the power of the normal test was considerably larger than the power of the exact test. Needless to say, the article by Berkson prompted several replies, most notably by Barnard [69], Basu [85], Corsten and de Kroon [286], and Kempthorne [722]; see also a 1984 article by Yates [1476, pp. 439–441].[23]

Permutations of cell frequencies for contingency tables in this period were not limited to determination of the exact Fisher–Yates probability value for 2 × 2 contingency tables. In 1970 Pierce developed an ALGOL computer program for computing the Fisher–Yates exact probability value for a 2 × 3 contingency table [1127, pp. 129–130, 283–287]. Pierce used a recursive procedure that essentially eliminated all factorial expressions and provided the opportunity to combine groups of constants for storage in the computer memory.

In 1972, March published algorithm CONP in FORTRAN for computing exact probabilities for $r \times c$ contingency tables. As March noted, if a sample of size N is subjected to two different and independent classifications, A and B, with R and C classes, respectively, the probability P_x of obtaining the observed array of cell frequencies $X(x_{ij})$, under the conditions imposed by the arrays of marginal frequency totals $A(r_i)$ and $B(c_j)$, is given by

$$P_x = \frac{\displaystyle\prod_{i=1}^{R} r_i! \prod_{j=1}^{C} c_j!}{\displaystyle N! \prod_{i=1}^{R} \prod_{j=1}^{C} x_{ij}!} .$$

[23] The controversy was to become a long-standing argument as to the proper method to analyze 2 × 2 contingency tables when both marginal frequency distributions were considered to be fixed, only one marginal frequency distribution was considered to be fixed, or neither marginal frequency distribution was considered to be fixed. In this regard, see also two articles by Barnard in 1947 [66, 67], an article by Plackett in 1977 [1137], an article by Yates in 1984 [1476], and an article by Campbell in 2007 [239].

The method utilized by March was to redefine P_x as

$$P_x = \frac{Q_x}{R_x} ,$$

where

$$Q_x = \frac{\prod_{i=1}^{R} r_i! \prod_{j=1}^{C} c_j!}{N!} ,$$

which, as March noted, was constant for the given set of marginal frequency totals, r_i and c_j, and

$$R_x = \prod_{i=1}^{R} \prod_{j=1}^{C} x_{ij}! ,$$

which varied depending on the array of cell frequencies (x_{ij}). March then used floating point logarithms (base 10) to compute the factorial expressions up to 100; above 100 he used Stirling's approximation. He tested the program using 2 × 3 contingency tables with $N = 30$, 2 × 4 contingency tables with $N = 7$, and 3 × 3 contingency tables with $N = 7$.

James Stirling

James Stirling was born in May 1692 in Garden, Stirlingshire, approximately 20 km from the town of Stirling, Scotland. Nothing is known of Stirling's early childhood, but it is documented that he enrolled in Balliol College, University of Oxford, in 1710 as a Snell Exhibitioner and was further awarded the Bishop Warner Exhibition scholarship in 1711.[24] Stirling lost his funding when, because he was a Jacobite, he refused to swear a loyalty oath to the British Crown. His refusal to swear the oath meant that Stirling could not graduate; however, he remained at Oxford for 6 years, until 1717 [1046].

In 1717 Stirling published his first paper extending a theory of plane curves by Newton, who was provided a copy of the paper. That same year, Stirling traveled to Venice where it is thought that he expected to become Chair of Mathematics, but for reasons unknown the appointment fell through. In 1722

(continued)

[24] At the University of Oxford, and other universities in England, an Exhibition is a financial grant or bursary awarded on the basis of merit. The recipient is an exhibitioner. The amount awarded is usually less than a Scholarship.

Stirling returned to Scotland and in late 1724 moved to London to teach mathematics at William Watt's Academy, in part on the recommendation of Newton. In 1730, Stirling produced *Methodus Differentialis sive Tractatus de Summatione et Interpolatione Serierum Infinitarum* (Differential Method with a Tract on Summation and Interpolation of Infinite Series) [1046]. It was Example 2 to Proposition 28 that became Stirling's most important and enduring work, his asymptotic calculation for $n!$ or "Stirling's approximation." Stirling's formula, given by

$$n! \doteq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \ ,$$

is in fact the first approximation to what is called "Stirling's series" given by

$$n! \doteq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51{,}840n^3} - \frac{571}{2{,}488{,}320n^4} + \frac{163{,}879}{209{,}018{,}880n^5} + \cdots\right)$$

[2, p. 257]. The asymptotic expansion of the natural logarithm of $n!$ is also referred to as "Stirling's series" and is given by

$$\ln n! \doteq n \ln n - n + \frac{1}{2}\ln(2\pi n) + \frac{1}{12n} - \frac{1}{360n^3} + \frac{1}{1{,}260n^5} - \frac{1}{1{,}680n^7} + \frac{1}{1{,}188n^9} - \frac{691}{360{,}360n^{11}} \cdots$$

[1173].[25] Sir Isaac Newton was elected President of the Royal Society in 1724 and served until his death in 1727. Stirling, on the recommendation of Newton, was elected Fellow of the Royal Society in 1726. James Stirling F.R.S. died in Edinburgh on 5 December 1770 at age 78.

The procedure by March [890] prompted several comments. Boulton noted that the method proposed by March was rather inefficient as it operated by generating all combinations that satisfied a weakened set of constraints, rejecting those combinations that violated the constraints imposed by the observed marginal frequency totals [185, p. 326]. Boulton modified the algorithm by March, utilizing

[25] It should be noted that in 1730 Abraham de Moivre published his *Miscellanea Analytica de Seriebus et Quadraturis* in which de Moivre first gave the expansion of factorials now known as Stirling's series, which should probably be referred to as the de Moivre–Stirling series [361, p. 8].

a procedure by Boulton and Wallace [187]. In 1975 Hancock [583] also revised the algorithm by March and compared the two methods. Hancock found his proposed modification to be much faster than the March algorithm. As Hancock illustrated, for a 4 × 4 contingency table with all cell frequencies equal to 1, the March algorithm examined 1,953,125 contingency tables before it reached a result, compared with only 10,147 contingency tables for Hancock's revised method. In 1976 Boulton also compared times for his 1973 algorithm with the algorithm of Hancock and found the Boulton procedure to be faster than the Hancock procedure [186].

In 1976 Howell and Gordon [657] published FORTRAN subroutine CONTIN, which was just a special case of the general method for an $r \times c$ contingency table with fixed marginal frequency totals described previously by Freeman and Halton in 1951 [480]. The Howell and Gordon procedure enumerated all possible $r \times c$ contingency tables, given fixed marginal frequency totals, and calculated the exact probability of the observed contingency table or one more extreme. The subroutine relied on the inefficient calculation of factorials and was based on the formula for the exact probability of a single $r \times c$ contingency table, given the marginal frequency distributions,

$$P_k = \frac{\displaystyle\prod_{i=1}^{r} R_{i.}! \prod_{j=1}^{c} C_{.j}!}{N! \displaystyle\prod_{i=1}^{r} \prod_{j=1}^{c} X_{ij}!} ,$$

where $R_{i.}$ and $C_{.j}$ denoted the fixed row and column marginal frequency totals, N was the total number of observations, X_{ij} denoted the observed cell frequencies, and P_k was the probability of contingency table k.

The year 1977 was to be important for calculating exact probability values for $r \times c$ contingency tables. First, Fleishman developed a program for calculating the exact probability for $r \times c$ contingency tables [466]. The program by Fleishman was based on an extension of the Fisher–Yates exact probability test and utilized the general method of Freeman and Halton [480] (q.v. page 172). As Fleishman noted, for a 4 × 3 contingency table with a total frequency of 82 there were over 804,000 contingency tables enumerated, and the execution time was over 40 min on an IBM 360/75, graphically illustrating the difficulties of computing at the time.

Second, Agresti and Wackerly published an article on exact conditional tests of independence for $r \times c$ contingency tables with fixed marginal frequency totals [7]. In this case, Agresti and Wackerly were less concerned with the exact hypergeometric probability and more concerned with the exact probability of established statistics, such as Pearson's chi-squared statistic. They were able to show that exact tests of independence using the chi-squared formula, or any measure of association as the test statistic, were manageable, i.e., required less than a minute of CPU times on an IBM 360/165 mainframe computer for a variety of 2 × 3, 2 × 4, 2 × 5, 2 × 6, 2 × 7, 3 × 3, 3 × 4, 3 × 5, and 4 × 4 contingency tables.

Agresti and Wackerly noted that procedures that ordered sample points solely on the basis of the probability of occurrence had received strong criticism, especially from Radlow and Alf in 1975 [1150]. The rationale of the criticism was that some configurations of cell frequencies may be less likely than the observed table under the null hypothesis, but exhibit less discrepancy from the null hypothesis than the observed table. Thus, Agresti and Wackerly defined the attained significance level to be the sum of the probability values of all tables for which the value of the test statistic was at least as large as the value of the test statistic for the observed table [7, p. 114]. This was destined to become an important observation.

Third, Bedeian and Armenakis developed a program for computing the Fisher–Yates exact probability test and the coefficient of association lambda (λ) for $r \times c$ contingency tables [91]. The stated purpose of the Bedeian and Armenakis paper was to provide a mathematical algorithm for the Fisher–Yates exact test that was adaptable to $r \times c$ contingency tables and also provided the user with λ, an index of predictive association designed for cross-tabulation of two nominal-level variables developed by Leo A. Goodman and William H. Kruskal at the University of Chicago in 1954 [534].[26]

Goodman–Kruskal's Lambda

Lambda was developed by Leo A. Goodman and William H. Kruskal in 1954 at the University of Chicago [534]. Lambda was designed to measure the degree of association between two categorical (nominal-level) variables that had been cross-classified into an $r \times c$ contingency table.

Three lambda coefficients were defined: two asymmetric and one symmetric. The first, $\lambda_{\text{rows}|\text{cols}}$, was for cases when the column variable was the independent variable and the row variable was the dependent variable; the second, $\lambda_{\text{cols}|\text{rows}}$, was for cases when the row variable was the independent variable and the column variable was the dependent variable; and the third, $\lambda_{\text{symmetric}}$, was essentially an average of $\lambda_{\text{rows}|\text{cols}}$ and $\lambda_{\text{cols}|\text{rows}}$.

Perfect association between the row and column variables results in $\lambda = 1$, and $\lambda = 1$ implies perfect association. On the other hand, independence of the row and column variables results in $\lambda = 0$, but $\lambda = 0$ does not necessarily imply independence. The reason for this is that λ is ultimately based on modal values and if the modal values of one variable all occur in the same category of the other variable, λ defaults to zero, a serious deficiency of Goodman and Kruskal's λ.

[26] The 1954 article by Goodman and Kruskal was the first of four articles on measures of association for cross classifications published in *Journal of the American Statistical Association* in 1954, 1959, 1963, and 1972 [534–537]. Robert Somers referred to the first of these papers as "a landmark to those working with statistics in the behavioral sciences" [1295, p. 804] and Stephen Fienberg was quoted as saying that the series constituted "four landmark papers on measures of association for cross classifications" [538, p. v].

Fig. 4.6 The 2×3 contingency table of Bedeian and Armenakis

2	2	142	146
0	2	106	108
2	4	248	254

The table used by Bedeian and Armenakis to illustrate their procedure is given in Fig. 4.6 [91, p. 256]. Based on the 2×3 contingency table in Fig. 4.6, Bedeian and Armenakis calculated the Fisher–Yates exact probability value to be 0.11944, and the lambda values to be $\lambda_{\text{rows}|\text{cols}} = 0.00$, $\lambda_{\text{cols}|\text{rows}} = 0.00$, and $\lambda_{\text{symmetric}} = 0.00$.

Unfortunately, 0.11944 is the point-probability value for the observed table in Fig. 4.6 rather than the two-tailed probability value.[27] For the table in Fig. 4.6, there are 15 possible cell configurations given the fixed marginal frequency totals, of which 13 are as or more extreme than the observed cell frequencies, resulting in an exact two-tailed probability value of 0.6636.

Even more unfortunate was the poor choice by Bedeian and Armenakis for the example 2×3 contingency table depicted in Fig. 4.6. Since the modal values of the row variable (142 and 106) both occur in the same category of the column variable (third column), and the modal values of the column variable (2, 2, and 142) all occur in the same category of the row variable (first row), then all three lambda coefficients are necessarily zero.

Finally in 1977, Baker introduced FORTRAN subroutine TABPDF that evaluated an $r \times c$ contingency table for three models [55]. The first model considered both sets of marginal frequency totals as fixed. The second model considered the row marginal frequency totals only as fixed, so that the summation of the probability values was over all $r \times c$ contingency tables with marginal frequency totals consistent with just the row marginal frequency totals. The third model considered neither row nor column marginal frequency totals as fixed, so that the summation of the probability values was over all $r \times c$ contingency tables with the same frequency total.[28]

4.6 Siegel–Tukey and Tables for the Test of Variability

To this point, Chap. 4 has primarily considered the decades of the 1960s and 1970s in terms of computing power, concentrating on the contributions of researchers who provided algorithms for generating random permutation sequences, computing exact and resampling-approximation probability values, the analysis of contingency tables, and a moment-approximation approach designed specifically for permutation tests, per se. The remainder of the chapter is dedicated to the statistical permutation

[27] The correct point-probability value for the table in Fig. 4.6 is 0.11946.

[28] The three research designs were first described by George Barnard in a 1947 *Biometrika* article [67] (q.v. page 130).

literature that was developed in the same period. It begins with a 1960 article on two-sample rank tests by Sidney Siegel and John Tukey.

Work on the publication of tables that listed exact probability values for a variety of rank tests that had begun in the 1950s continued in the 1960s. In 1960 Siegel and Tukey developed a non-parametric two-sample test based on differences in variability between the two unpaired samples, rather than the more conventional tests for differences in location [1273]. The Siegel–Tukey test was designed to replace parametric F tests for differences in variances that depended heavily on normality, such as Bartlett's F and Hartley's $F_{\max}$ tests for homogeneity of variance [78,596]. Within this article Siegel and Tukey provided tables of one- and two-sided critical values based on exact probabilities for a number of levels of significance.[29]

S. Siegel

Sidney Siegel received his B.A. degree from San José State College (now, San José State University) in 1951 and his Ph.D. in psychology from Stanford University in 1953 [1271]. It was while Siegel was a graduate student at Stanford that he was first exposed to statistics, studying under Quinn McNemar, Lincoln Moses, George Polya, Albert Bowker, Kenneth Arrow, and John Charles Chenoweth (J.C.C.) McKinsey. He served for 1 year as a Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford, thereafter he was employed at Pennsylvania State University. He was the author of *Nonparametric Statistics for the Behavioral Sciences*, which ultimately became one of the best selling statistics books of all time, appearing in English, Japanese, Italian, German, and Spanish [1272]. Sidney Siegel passed away on 29 November 1961 at the early age of 45 from coronary thrombosis [1271, p. 16]. His book was resurrected and revised in 1988 by N. John Castellan and published as a second edition with authors Siegel and Castellan. N. John Castellan died at home on 21 December 1993 at the age of 54.

J.W. Tukey

John Wilder Tukey received his B.A. and M.A. degrees in chemistry from Brown University in 1936 and 1937, respectively, and his Ph.D. in mathematics from Princeton University in 1939 under the supervision of the algebraic topologist Solomon Lefschetz, followed by an immediate appointment as

(continued)

[29] In 1960 A.R. Ansari and R.A. Bradley published an article titled "Rank sum tests for dispersion" that provided tables of critical values for the symmetrical version of the Siegel–Tukey test and also discussed the normal approximation to the null distribution [812, p. 52].

Henry B. Fine Instructor in Mathematics [425]. A decade later, at age 35, he was advanced to Professor and in 1976 he was awarded the Donner Professor of Science chair. Tukey spent his entire academic career at Princeton University, but simultaneously worked for 40 years in the Department of Statistics and Data Analysis at the AT&T Bell Laboratories (now, Alcatel–Lucent) in Murray Hill, New Jersey, until his retirement in 1985. In 1956 Tukey assumed the directorship of the newly founded Statistical Research Group at Princeton and then Head of the Department of Statistics at Princeton when it was established in 1965. In 1973, President Nixon awarded Tukey the National Medal of Science.

Among his many accomplishments, Tukey is known for his work on exploratory data analysis (EDA), his coining of the word "software" in the January 1958 issue of *American Mathematical Monthly* [1256, p. 772], and his invention of the word "bit" to represent a binary digit in 1946.[30] His collaboration with fellow mathematician James William Cooley resulted in the discovery of the fast Fourier transform (FFT), which was to become important in permutation methods in the 1990s. Tukey held honorary degrees from the Case Institute of Technology, the University of Chicago, and Brown, Temple, Yale, and Waterloo Universities; in June 1998, he was awarded an honorary degree from Princeton University [425]. The eight volumes of *The Collected Works of John W. Tukey* provide an excellent compendium of the writings of John Tukey, as well as a rich source of biographical material [207, 212, 213, 258, 294, 700, 701, 871]. John Wilder Tukey passed away from a heart attack that followed a brief illness on 26 July 2000 at the age of 85 [214, 704].

Let the two sample sizes be denoted by n and m with $n \leq m$ and assign ranks to the $n+m$ ordered observations with low ranks assigned to extreme observations and high ranks assigned to central observations. More specifically, assign rank 1 to the smallest value, rank 2 to the largest value, rank 3 to the second largest value, rank 4 to the second smallest value, rank 5 to the third smallest value, and so on, alternately assigning ranks to the end values two at a time (after the first) and proceeding toward the middle. Since the sum of the ranks is fixed, Siegel and Tukey chose to work with the sum of ranks for the smaller of the two samples, represented by R_n. They also provided a table with one- and two-sided critical values of R_n for $n \leq m \leq 20$ for various levels of α.

[30] The first use of the acronym "bit" for "binary digit" is often attributed to Claude Elwood Shannon of Bell Laboratories, the father of information science, as it was contained in his 1948 paper on "A mathematical theory of communication," e.g., [1166, p. 199]. However, in this seminal paper Shannon gave full credit to John Tukey for first suggesting the term [1254].

Fig. 4.7 Observations and dispersion ranks for ten graduate students

$\underline{20}$	$\underline{22}$	$\underline{23}$	24	25	27	$\underline{28}$	30	$\underline{32}$	45
1	4	5	8	9	10	7	6	3	2

For an example, suppose that there are $n = 5$ male and $m = 5$ female students in a graduate seminar and the observations are the ages of $n + m = 10$ students, where the ages of the male graduate students are $\{20, 22, 23, 28, 32\}$ and the ages of the female graduate students are $\{24, 25, 27, 30, 45\}$. The dispersion ranks are depicted in Fig. 4.7 where the ages of the male graduate students are underlined and $R_n = 1 + 4 + 5 + 7 + 3 = 20$.

A serious problem with the Siegel–Tukey test is its lack of symmetry. Another test with exactly the same properties can be obtained by reversing the pattern of Fig. 4.7, assigning rank 1 to the largest observation, rank 2 to the smallest, and so on [812, p. 33]. For the data in Fig. 4.7 this would yield $R_n = 2 + 3 + 6 + 8 + 4 = 23$. In 1962 Klotz [759] demonstrated the equivalence of the Siegel–Tukey test and comparable tests by Barton and David [82] and Freund and Ansari [481].

Siegel and Tukey noted that their choice of ranking procedure, with low ranks assigned to extreme observations and high ranks assigned to central observations, allowed the use of the same tables as were used for the Wilcoxon two-sample rank-sum test for location [1453]. Thus, they explained, their new test might "be considered a Wilcoxon test for spread in unpaired samples" [1273, p. 431]. Alternatively, as they explained, the Siegel–Tukey tables were equally applicable to the Wilcoxon, Mann–Whitney, White, and Festinger rank-sum procedures for relative location of two independent samples [427, 880, 1441, 1453], and were appropriate linear transformations of the tabled values published by Auble in 1953 [40].[31]

4.7 Other Tables of Critical Values

In 1961 Glasser and Winter published a paper containing approximate critical values for Spearman's rank-order correlation coefficient, r_s, for one-tailed α levels of 0.001, 0.005, 0.010, 0.025, 0.050, and 0.100 with $n = 11, 12, \ldots, 30$ for use in testing the null hypothesis of independence [516]. Noting that exact probability values for r_s had been calculated for samples up to size $n = 10$ by Olds in 1938 [1054], Kendall, Kendall, and Babington Smith in 1939 [746], and David, Kendall, and Stuart in 1951 [328], Glasser and Winter used a Gram–Charlier Type A series approximation to the distribution function of r_s first given by David, Kendall, and Stuart in 1951 [328] to extend the tables of r_s to $n = 30$.

[31] Siegel and Tukey did not mention, and were apparently unaware of, the equivalent tests by J.B.S. Haldane and C.A.B. Smith, published in 1948 in *Annals of Genetics* (q.v. page 154), and by D. van der Reyden, published in 1952 in *Rhodesia Agricultural Journal* (q.v. page 165).

In 1963 Verdooren published new tables of exact critical values for the Wilcoxon–Mann–Whitney two-sample rank-sum sample statistic for lower significance levels 0.001, 0.005, 0.01, 0.025, 0.05, and 0.10 for sample sizes $m \leq n = 1, 2, \ldots, 25$ [1398]. In an appendix to the article, Verdooren listed errata correcting a few of the values provided in the earlier tables published by White in 1952 [1441], Auble in 1953 [40], and Siegel and Tukey in 1960 [1273]. Also in 1963, Bennett and Nakamura published tables for testing significance in 2×3 contingency tables [94]. If A_j for $j = 1, 2, 3$ denotes the three column marginal frequency totals and $A_1 = A_2 = A_3$, then four significance levels were tabulated by Bennett and Nakamura using the randomized test principle of Freeman and Halton: 0.05, 0.025, 0.01, and 0.001.

In 1964 Milton published a new table of critical values for the Wilcoxon–Mann–Whitney two-sample rank-sum sample statistic [996], extending previous tables published by Wilcoxon [1454], White [1441], van der Reyden [1391], Auble [40], Siegel [1272], Rümke and van Eeden [1206], Jacobson [677], Verdooren [1398], and Owen [1075]. The extended tables were for one-tailed α levels of 0.0005, 0.0025, 0.005, 0.001, 0.01, 0.025, 0.05, 0.10, and for sample sizes of $n \leq 20$ and $m \leq 40$.

4.8 Edgington and Randomization Tests

Beginning in the early 1960s, Eugene Edgington at the University of Calgary published a number of books and articles on permutation methods and was an influential voice in promoting the use of permutation tests and measures, especially to psychologists and other social scientists. Edgington was especially critical of the use of normal-theory methods when applied to nonrandom samples.

E.S. Edgington

Eugene S. Edgington, "Rusty" to his friends, received his B.S. and M.S. degrees in psychology from Kansas State University in 1950 and 1951, respectively, and his Ph.D. in psychology from Michigan State University in 1955. He has enjoyed a long career in the Department of Psychology at the University of Calgary, Alberta, where he is now Emeritus Professor. Edgington has published many books and articles dealing with permutation methods, the best known of which is *Randomization Tests*, first published in 1980 and continued through four editions, the last co-authored with Patrick Onghena at the Katholieke Universiteit, Leuven, in 2007. Edgington has been instrumental in the development of permutation tests for experimental designs and an influential voice in the promotion of permutation methods, especially among psychologists. Eugene S. Edgington is presently Professor Emeritus at the University of Calgary, Alberta.

In 1964 Edgington published a descriptive article on randomization tests in *The Journal of Psychology* [387]. This article marked the beginning of an important series of articles and books on methods and introduced permutation statistics to a wide audience of psychologists, who subsequently found permutation methods both popular and useful.[32,33] In this brief article, Edgington defined a randomization test as a statistical test that derives a sampling distribution of a statistic from repeated computations of the statistic for various ways of pairing or dividing the scores [387, p. 445]. He considered three types of randomization tests: tests for differences between independent samples, tests for differences between paired samples, and tests of correlation. Edgington noted that randomization tests could be particularly useful whenever the assumptions of parametric tests could not be met, when samples were very small, and when probability tables for the desired statistic were not available.

In 1966 Edgington published an article on statistical inference and nonrandom samples [388]. Writing primarily for psychologists, Edgington pointed out that since experimental psychologists seldom sample randomly, it was difficult for psychologists to justify using hypothesis-testing procedures that required the assumption of random sampling of the population or populations about which inferences were to be made. Edgington stated his position unequivocally, "statistical inferences cannot be made concerning populations that have not been randomly sampled" [388, p. 485].[34] In a concession to psychological researchers, however, he also pointed out that non-statistical inferences could, of course, be drawn on the basis of logical considerations. He went on to advocate the use of permutation methods for statistical inferences from nonrandom samples, but also stated that "this does not imply that parametric tests cannot be used" [388, p. 487]. He explained that a researcher could use parametric tests as approximations to permutation tests, echoing previous studies by Silvey [1276], Wald and Wolfowitz [1407], Friedman [486], Eden and Yates [379], Kempthorne [718, p. 152], Pitman [1131], and Welch [1428], supporting the claim that permutation tests are the gold standard against which parametric tests are to be evaluated.

In a 1967 article on making statistical inferences from a sample of $n = 1$, Edgington further clarified the problem of making statistical inferences with permutation methods [389]. He noted that while it was certainly correct that a researcher could not statistically generalize to a population from only one subject, it was also correct that a researcher could not statistically generalize to a population from which the researcher had not taken a random sample of subjects. He noted that this observation ruled out making inferences to populations for virtually all

[32] Authors' note: Edgington termed exact permutation tests "randomization tests" and resampling permutation tests "approximate randomization tests"; we follow his convention in this section.

[33] Psychologists typically study small nonrandom samples, for which permutation tests are ideally suited (q.v. page 274).

[34] See also a stern warning about the use of convenience samples in research in a 1991 textbook on *Statistics* by Freedman, Pisani, Purves, and Adhikari [479, p. 506].

psychological experiments, both those with large and small sample sizes.[35] Finally, he noted that hypothesis testing was still possible without random samples, but that significance statements were consequently limited to the effect of the experimental treatment on the subjects actually used in the experiment, with generalization to other subjects being based on logical, non-statistical considerations [389, p. 195].

In 1969, in the same journal, Edgington elaborated on approximate randomization tests, i.e., resampling or Monte Carlo permutation tests, which he had touched on only briefly in his 1964 paper [390]. He defined an approximate randomization test as a test in which the significance of an obtained statistic was determined by using an approximate sampling distribution consisting of a random sample of statistics randomly drawn from the entire sampling distribution [390, p. 148]. An approximate randomization test could thereby greatly reduce the amount of computation to a practical level. As an example, Edgington considered an approximate randomization test on a correlation coefficient with $n = 13$ subjects, yielding $13! = 6{,}227{,}020{,}800$ equally-probable pairings. He estimated the full randomization test would take 197 years of continuous 24-h-a-day operation to compute all the correlation coefficients in the entire sampling distribution [390, p. 144].

In an important statement, Edgington argued that in an approximate randomization test, the significance of an obtained statistic was determined by reference to a distribution "composed of the approximate sampling distribution *plus* the obtained statistic." The significance was the proportion of statistics within this distribution that were as large or larger than the obtained statistic [390, p. 148]. This was an important observation at the time as Edgington and others recommended computing only 999 statistics plus the obtained statistic. Today, with 1,000,000 sample statistics being regularly generated, it is perhaps a moot point.

Also in 1969, Edgington published a book on *Statistical Inference: The Distribution-free Approach* that contained an entire chapter on randomization tests for experiments [391]. In this lengthy 76-page chapter, Edgington examined inferences concerning hypotheses about experimental treatment effects with finite populations, with no assumptions about the shapes of the populations, and for nonrandom samples. He explored in great detail and with many examples, randomization tests for paired comparisons, contingency tables, correlation, interactions, differences between independent samples, and other randomization tests such as differences between medians, ranges, and standard deviations. Near the end of the chapter, Edgington turned his attention once again to approximate randomization tests (resampling-approximation permutation tests), noting that the amount of computation for a randomization test could be reduced to a manageable level by using random samples of all the pairings of divisions in the entire sampling distribution to obtain a smaller sampling distribution [391, p. 152].

[35] This was very much the same conclusion that John Ludbrook and Hugh Dudley came to regarding biomedical research in an article on "Why permutation tests are superior to t and F tests in biomedical research," published in *The American Statistician* in 1998. [856].

Edgington concluded the chapter with a discussion of normal-theory tests as approximations to randomization tests. He argued that when an experiment has been designed so that a randomization test can be carried out, a normal-theory test can sometimes be used as an approximation to the randomization test. His logic, as he explained, was that normal-theory tests can be regarded as approximations to randomization tests to the extent that the sampling distribution of the relevant statistic, such as a t or F, underlying the probability tables is similar to that for a randomization test using the same statistic [391, p. 161].

In 1973 Edgington and Strain summarized computer time requirements for a number of statistical tests, both exact randomization tests and approximate randomization tests [397]. Included were exact and approximate randomization t tests for two independent samples, matched-pairs, one-way analysis of variance, and randomized blocks analysis of variance. All the tests were conducted on a CDC 6400, which was a state of the art mainframe computer at that time. In this article, Edgington and Strain made the observation that when subjects have not been randomly selected from a defined population, but have been randomly assigned to treatments, randomization tests are the only valid tests that can be performed [397, p. 89].[36] On this topic, see also a 1972 article by Youden [1478] and a 1977 article by Kempthorne [721].

On this topic, Box, Hunter, and Hunter noted that the randomization tests introduced by Fisher in 1935 and 1936 were early examples of what were to be later called, in spite of Fisher's protest, "nonparametric" or "distribution-free" tests. They argued that (1) unless randomization has been performed, then "distribution-free" tests do not possess the properties claimed for them, and (2) if randomization has been performed, standard parametric tests usually supply adequate approximations [194, p. 104].

4.9 The Matrix Occupancy Problem

The matrix occupancy problem, as discussed by Mielke and Siddiqui in 1965 [988], was motivated by a study of bronchial asthma associated with air pollutants from the grain mill industry in Minneapolis, Minnesota, in 1963 [291]. Following the notation of Mielke and Siddiqui [988], consider a $b \times g$ occupancy matrix with $b \geq 2$ asthmatic patients (rows) and $g \geq 2$ days (columns). Let X_{ij} denote the observation of the ith patient, $i = 1, \ldots, b$, on the jth day, $j = 1, \ldots, g$, where $X_{ij} = 1$ if an asthmatic attack occurred and $X_{ij} = 0$ if no asthmatic attack occurred.

[36] This is an important point emphasized by others; viz., permutation tests on experiments are valid only when preceded by randomization of treatments to subjects; see also a 1963 article by Cox and Kempthorne [293, p. 308] and a 1988 paper by Tukey [1382]. The randomization of treatments to subjects was one of the few points on which Neyman and Fisher agreed, as both felt that it provided the only reliable basis for dependable statistical inference [816, p. 76].

Mielke and Siddiqui presented an exact permutation procedure for the matrix occupancy problem that is most appropriate for small b and large g [988]; see also a discussion by Mielke and Berry in 2007 [965, pp. 135–138]. Let

$$R_i = \sum_{j=1}^{g} X_{ij}$$

be a fixed row total, $i = 1, \ldots, b$, let

$$M = \prod_{i=1}^{b} \binom{g}{R_i}$$

denote the total number of equally-likely distinguishable $b \times g$ occupancy matrices under the null hypothesis, and let $w = \min(R_1, \ldots, R_b)$. If U_k is the number of distinct $b \times g$ matrices with exactly k columns filled with 1s, then

$$U_w = \binom{g}{w} \prod_{i=1}^{b} \binom{g-w}{R_i - w}$$

is the initial value of the recursive relation

$$U_k = \binom{g}{k} \left[\prod_{i=1}^{b} \binom{g-k}{R_i - k} - \sum_{j=k+1}^{w} \binom{g-k}{j-k} \frac{U_j}{\binom{g}{j}} \right] ,$$

where $0 \leq k \leq w - 1$. If $k = 0$, then

$$M = \sum_{k=0}^{w} U_k$$

and the exact probability value under the null hypothesis of observing s or more columns exactly filled with 1s is

$$P_s = \frac{1}{M} \sum_{k=s}^{w} U_k \ , \tag{4.1}$$

where $0 \leq s \leq w$. Eicker, Siddiqui, and Mielke described extensions to the matrix occupancy problem solution in 1972 [405].

For an example, consider an experiment with $b = 6$ asthmatic patients examined over a series of $g = 8$ days. The data are summarized in Table 4.3. The R_i asthmatic patient attack totals are $\{4, 6, 5, 7, 4, 6\}$; the minimum of R_i, $i = 1, \ldots, b$, is $w = 4$; the number of observed days filled with 1s is $s = 2$ (days 2 and 7);

$$\sum_{k=s}^{w} U_k = \sum_{k=2}^{4} U_k = 149{,}341{,}920 + 6{,}838{,}720 + 40{,}320 = 156{,}220{,}960 \ ;$$

Table 4.3 Attack (1) and no attack (0) for $b = 6$ asthmatic patients over a series of $g = 8$ days

	Days								
Patient	1	2	3	4	5	6	7	8	R_i
1	0	1	1	1	0	0	1	0	4
2	1	1	1	0	0	1	1	1	6
3	0	1	0	1	1	0	1	1	5
4	1	1	1	1	0	1	1	1	7
5	0	1	1	0	0	0	1	1	4
6	1	1	1	1	0	1	1	0	6

$M = 1{,}721{,}036{,}800$; and following Eq. (4.1),

$$P_s = \frac{156{,}220{,}960}{1{,}721{,}036{,}800} = 0.0908 \; .$$

For a series of articles published in *The American Statistician*, Nathan Mantel [882] observed in 1974 that the solution to the matrix occupancy problem described by Mielke and Siddiqui [988] was also the solution to the "committee problem" considered by Mantel and Pasternack in 1968 [888], Gittelsohn in 1969 [514], Sprott in 1969 [1314], and White in 1971 [1442]. Specifically, the committee problem involves $g \geq 2$ members, $b \geq 2$ committees, and $X_{ij} = 0$ if the jth member, $j = 1, \ldots, g$, belongs to the ith committee, $i = 1, \ldots, b$, and $X_{ij} = 1$ if the jth member does not belong to the ith committee. Under the null hypothesis that all distinguishable $b \times g$ matrices are equally likely, the exact probability value of observing s columns exactly filled with 1s is the previously-defined P_s in Eq. (4.1). Thus, in 1974 the matrix occupancy and committee problems were conclusively shown to be identical by Mantel [882].

4.10 Kempthorne and Experimental Inference

In this period, no one did more to promote permutation methods, advocate their use over parametric methods, and extol their virtues than Eugene Edgington of the University of Calgary, John Tukey of Princeton University, Alvan Feinstein of the Yale University School of Medicine, and Oscar Kempthorne of Iowa State University. Together, their influential voices advocated the use of permutation methods to the exclusion of classical methods. The work of Kempthorne, in particular, has not been received with the gravitas it deserves. In 1963 Kempthorne published an article with David Cox on permutation methods for comparing survival curves [293] and in 1966 Kempthorne published an article on experimental inference [720]. The 1966 article was the text of the Fisher Memorial Lecture given by Kempthorne in Philadelphia on 10 September 1965 to a joint session of the American Statistical Association, the Institute of Mathematical Statistics, and the Biometrics Society.[37]

O. Kempthorne

Oscar Kempthorne earned his B.A. and M.A. degrees from Clare College, University of Cambridge, in 1940 and 1943, respectively, and an honorary Sc.D. from the University of Cambridge in 1960. He was employed at the Rothamsted Experimental Station from 1940 to 1946, where he worked with both R.A. Fisher and Frank Yates. In 1947 Kempthorne accepted a position as Associate Professor at Iowa State College (now, Iowa State University) in Ames, Iowa. He was promoted to Professor in 1951 and was named Distinguished Professor in Sciences and Humanities at Iowa State University in 1964 [474, 624].

Kempthorne is considered the founder of the "Iowa school" of experimental design and analysis of variance. His contributions centered largely on three major areas: experimental design, genetic statistics, and the philosophy of statistics. In the present context, Kempthorne's many contributions to randomization theory is of primary importance. Kempthorne was highly critical of both model-based inference and Bayesian statistics and strongly embraced a randomization approach to the statistical analysis of experiments. Much of the work he did on randomization theory is summarized in his 1994 book on *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design*, co-authored with Klaus Hinkelmann [625]. Oscar Kempthorne retired from Iowa State University in 1989 after an academic career of 42 years and passed away on 15 November 2000 in Annapolis, Maryland, at the age of 81 [626].

In the 1963 paper, Cox and Kempthorne, both at Iowa State University in Ames, Iowa, analyzed data from swine concerning the genetic effects of paternal irradiation on survival in the first generation [293]. Permutation test procedures were utilized for the evaluation of the results "because of the inherently complex correlational structure of the data" [293, p. 307]. Specifically, Cox and Kempthorne argued that a permutation test procedure has an essential role with vector observations whose correlational structure is complex and not estimable with any reasonable precision. The permutation method is stated so succinctly and concisely that it is worth summarizing here.

[37] There are several lecture series in honor of R.A. Fisher. This series is the distinguished R.A. Fisher Lectureship and Kempthorne was, in 1965, the second lecturer. The first in a long list of distinguished lecturers was Maurice Bartlett in 1964, and the third was John Tukey in 1966. The lecture series continues to this day.

Consider a permutation test to compare two groups. The permutation test procedure considers all possible partitions of the experimental units in two groups of the same size as those in the original experiment. A criterion of interest is evaluated for each partition and the proportion of the partitions where the value of the criterion equals or exceeds that observed in the original experiment is determined [293, p. 308].

Cox and Kempthorne emphasized that the significance of the partition of the experimental units observed in the actual experiment was to be judged directly from the proportion of the random partitions that yielded equal or greater values for the criterion under study. Then, following Fisher, they underscored that for the evaluation of significance to have validity, the original division of the experimental units "into groups subject to different treatments *must be done at random*" [293, p. 308].[38] Finally, Cox and Kempthorne noted that if the number of experimental units was large, an evaluation of all possible partitions became impractical and the solution was to examine a random sample of the possible partitions of the experimental units.

The actual experiment analyzed by Cox and Kempthorne consisted of 3,552 swine born into $N = 362$ litters, with $n_1 = 180$ from irradiated males and $n_2 = 182$ from control males. Since there were

$$\frac{N!}{n_1!\,n_2!} = \frac{362!}{180!\;182!} = 3.92 \times 10^{107}$$

possible partitions, a random sample of all possible partitions was extracted and only 200 partitions were examined. Three different permutation analyses were conducted with the resulting conclusion that paternal irradiation adversely affected early survival in swine.

In the 1966 paper Kempthorne treated permutation tests as a gold standard against which parametric tests might be evaluated: "[i]t may well be that the randomization [permutation] test is reasonably approximated by the t-test ..." [720, p. 20]. In addition, Kempthorne argued that the proper way to make tests of significance in simple randomized experiments is by a permutation test, concluding that "in the randomized experiment one should, logically, make tests of significance by way of the randomization test" [720, p. 21]. Much of the rest of the paper was devoted to illustrating permutation tests based on results obtained by T.E. Doerfler (a former student) and Kempthorne from a paired experiment, later published in 1969, where comparisons were made among the F test, the corresponding randomization test, the Wilcoxon matched-pairs rank-sum test, and the sign test [725].

[38] Emphasis added.

4.11 Baker–Collier and Permutation *F* Tests

In 1966 Frank B. Baker and Raymond O. Collier published three articles on permutation methods. The first of the three articles was a computer program for analysis of variance F tests by means of permutation [51]. The second article compared normal-theory and permutation variance-ratio tests on size and power of the appropriate F test for one-way and two-way completely randomized experimental designs [52]. The third article compared normal-theory and permutation variance-ratio tests on size and power of the appropriate F test for randomized blocks experimental designs [268].

4.11.1 A Permutation Computer Program

In 1966 Baker and Collier made available a FORTRAN computer program for calculating analysis of variance F tests by means of permutation [51]. The program was designed to handle any balanced fully-replicated or nested design with up to eight factors involving an equal number of observations per cell. The number of permuted samples was restricted to 1,000, as this consumed 4 min of run time on a CDC 6400 mainframe computer in 1966.

4.11.2 Simple Randomized Designs

Also in 1966 Baker and Collier considered "some empirical results on variance ratios under permutation" in the completely randomized analysis of variance design [52]. In general, they found high agreement between normal-theory and permutation tests of the probability of a type I error. They also found high agreement for power for variance-ratio tests in completely randomized designs. They concluded that the empirical results indicated that for the completely randomized design, the agreement between normal-theory F tests and permutation F tests for either size or power was acceptable over a range of skewness and kurtosis, and only extremely leptokurtic data affected the agreement, which they found to be mitigated by increasing sample size.

The empirical results were based on 1,000 random permutations of a basic data set utilizing a general purpose Monte Carlo computer routine due to Baker and Collier [51]. For a one-factor completely randomized design, Baker and Collier investigated two treatment layouts: one with two levels and one with three levels; in both cases they used six observations in each level, yielding a total number of observations of $N = 12$ and $N = 18$, respectively. They examined type I error at $\alpha = 0.10, 0.05, 0.025,$ and 0.01, with skewness values $g_1 = 0.00$ and $+1.00$, and kurtosis values $g_2 = -1.00, 0.00,$ and $+1.00$. For a two-factor completely randomized design they again investigated two treatment layouts: one a 2 × 2 factorial design and one a 3 × 2 factorial design, again with six subjects per cell for a total number of observations of $N = 24$ and $N = 36$, respectively. As in the one-factor design, they used $\alpha = 0.10, 0.05, 0.025$, and 0.01, $g_1 = 0.00$ and $+1.00$, and $g_2 = -1.00, 0.00$, and $+1.00$.

Baker and Collier concluded that their empirical results found high agreement between the variance-ratio test under permutation and under normal theory for the completely randomized analysis of variance design for a broad set of conditions. More specifically, the agreement was very good over all four treatment layouts when the skewness coefficients of the basic data were 0.00 and +1.00 and the kurtosis coefficients were −1.00, 0.00, and +1.00.

Finally, Baker and Collier noted that their investigations into the empirical determinations of the probability of type I error were in agreement with those given previously by Hack in 1958 [566], but that they had also extended Hack's results by analyzing both one- and two-way treatment layouts, and by studying data sets containing fewer observations than had Hack. In addition, their power estimates provided an extension of the randomized block results due to Kempthorne, Zyskind, Addelman, Throckmorton, and White to the completely randomized design published in 1961 [726].

4.11.3 Randomized Block Designs

While Baker and Collier [52] compared normal-theory and permutation tests on size and power for variance-ratio tests in completely randomized designs, Collier and Baker extended this work to simple randomized block designs [268]. Their investigation into randomized block designs was an attempt to supplement previous work on power by Kempthorne, Zyskind, Addelman, Throckmorton, and White in 1961 [726].

Because in an I treatments by J blocks design there are $(I\,!)^J$ possible permutations of the observed data, Collier and Baker chose to randomly sample permutations from four data sets, which they termed "basal responses," indicating responses that a set of hypothetical experimental units would produce under null treatment effects [268, p. 199]. To this end they investigated size and power under permutation of the usual F test of null treatment effects for two randomized block designs with one observation per cell: an $I = 3$ treatments by $J = 8$ blocks design and an $I = 3$ treatments by $J = 15$ blocks design.

Specifically, 24 observations for the 3 × 8 design and 45 observations for the 3 × 15 design were randomly sampled from each of (1) a normal distribution, (2) a log-normal distribution, (3) a one-sided exponential distribution, and (4) a two-sided exponential distribution. Analyses of the four sets of basal responses were based on Monte Carlo random permutations of 1,000 observations drawn from each data set.

For the power analyses, they examined power for the F test under permutation and under normal theory for both the 3 × 8 and 3 × 15 designs. For the 3 × 8 design they set normal-theory size to $\alpha = 0.05$ with normal-theory power of $\phi^2 = 0.80$ and 0.60, and $\alpha = 0.01$ with normal-theory power of $\phi^2 = 0.53$ and 0.32. For the 3 × 15 design they set normal-theory size to $\alpha = 0.05$ with normal-theory power of $\phi^2 = 0.80$ and 0.60, and $\alpha = 0.01$ with normal-theory power of $\phi^2 = 0.54$ and 0.32.

Collier and Baker found that under null treatment effects, the permutation distributions of F for the four sets of basal responses agreed quite well with the F distribution under normal theory for both the 3×8 and 3×15 factorial designs. Comparing the size of the F test under permutation with that expected from normal theory, they observed agreement of the permutation distribution in those regions "where significance levels are ordinarily set" [268, p. 203], and found little difference in results for the 3×8 and 3×15 designs.

Finally, they observed that the power of the F test under permutation compared favorably with the normal-theory counterpart. Taking the viewpoint that the value of the power of the F test under permutation represents an exact determination, Collier and Baker showed that the power of the test was slightly overestimated by a normal-theory power evaluation.

4.12 Permutation Tests in the 1970s

The development of permutation statistical methods expanded greatly in the 1970s. Among the many contributors was Alvan R. Feinstein who promoted permutation methods to clinical researchers in the early 1970s with a long series of articles published in *Clinical Pharmacology and Therapeutics* [e.g., 421]. In 1973 Dinneen and Blakesley published an algorithm for the Mann–Whitney U statistic [351]. In 1975 Arbuckle and Aiken published a program for Pitman's two-sample test [30], Patil published a program for Cochran's Q test [1090, p. 186], and Radlow and Alf published a method for computing an exact chi-squared test [1150]. In 1976 Mielke, Berry, and Johnson introduced multi-response permutation procedures (MRPP) that were designed especially for data-dependent permutation methods and relied on ordinary Euclidean distances instead of squared Euclidean distances [971]. In 1977 Gail and Mantel published an important technique whereby the number of possible arrangements of cell frequencies in $r \times c$ contingency tables could easily be estimated, given fixed marginal frequency totals [490]. In 1977 Soms published an algorithm for the Fisher–Pitman two-sample permutation test for differences between two independent samples [1296], Baker and Hubert published an article on inference procedures for ordering theory [53], and Green published a computer program for one- and two-sample permutation tests of location [548]. In 1979 Agresti, Wackerly, and Boyett developed a permutation procedure to provide resampling-approximation permutation tests for $r \times c$ contingency tables [8].

4.13 Feinstein and Randomization

Any chronicle of permutation methods would be incomplete without mention of Alvan R. Feinstein: mathematician, statistician, medical doctor, and founder of clinical epidemiology.

A.R. Feinstein

Alvan R. Feinstein earned his B.Sc. and M.Sc. degrees in mathematics at the University of Chicago in 1947 and 1948, respectively, and his M.D. degree at the University of Chicago School of Medicine in 1952. After completion of his residency in internal medicine at Yale University and Columbia–Presbyterian Hospital in New York, and a research fellowship at the Rockefeller Institute, he assumed the post of Medical Director of the Irvington House Institute (now part of New York University Langone Medical Center) in 1955. In 1962 Feinstein joined the faculty at the Yale University School of Medicine and in 1974 he became the founding director of the Robert Wood Johnson Clinical Scholars Program at Yale University [1026, 1310].

Feinstein is widely regarded as the founder of clinical epidemiology and patient-oriented medicine and the originator of clinimetrics: the application of mathematics to the field of medicine. Over his career, Feinstein published over 400 original articles and six books: *Clinical Judgment*, *Clinical Epidemiology*, *Clinimetrics*, *Clinical Biostatistics*, *Multivariate Analysis*, and *Principles of Medical Statistics*. At the time of his death of an apparent heart attack on 25 October 2001 at the age of 75, Alvan R. Feinstein was Sterling Professor of Medicine and Epidemiology, Yale University's most prestigious professorship, a position he occupied for the 10 years prior to his death [918, 1045].

In 1973 Feinstein published an article on "The role of randomization in sampling, testing, allocation, and credulous idolatry" [421]. The importance of this article was not that it contained new permutation methods, but that it summarized and promoted permutation methods to a new audience of clinical researchers in a cogent and lucid manner.[39,40] Feinstein, writing for a statistically unsophisticated readership, distinguished between socio-political research where the purpose was usually to estimate a population parameter, and medical research where the purpose was typically to contrast a difference between two groups. He observed that a random sample is mandatory for estimating a population parameter, but "has not been regarded as equally imperative for contrasting a difference" [421, p. 899]. As his focus was on medical investigations, he listed the major violations of the assumptions underlying tests of two groups:

[39] The 1973 Feinstein article was the 23rd in a series of informative summary articles on statistical methods for clinical researchers published in *Clinical Pharmacology and Therapeutics*. A collection of 29 of the articles written by Feinstein is available in *Clinical Biostatistics* where this article was retitled "Permutation tests and 'statistical significance'" [422].

[40] Authors' note: after 40-plus years, this 1973 article by Feinstein remains as perhaps the clearest non-mathematical introduction to permutation tests ever written and should be consulted by all researchers new to the field of permutation methods.

1. The groups studied in modern clinical or epidemiologic research are seldom selected as random samples.
2. For the many clinical and epidemiologic research projects that are performed as surveys, the subjects are not assigned randomly.
3. The distribution of the target variable is usually unknown in the parent population.
4. It is usually known that the target variable does not have a Gaussian distribution, and often departs from it dramatically.
5. It is usually known that the variances of the two samples are not remotely similar.

Feinstein then compared, in meticulous detail, the classical approaches embodied in the two-sample t test and the chi-squared test of independence for 2×2 contingency tables. For his example data, he noted that the probability values obtained from the classical approach differed substantially from those obtained from the corresponding permutation tests.[41] Regarding the chi-squared test of independence, Feinstein observed that the corresponding permutation test provided an exact answer to the research question that was "precise, unambiguous, unencumbered by any peculiar expectations about fractional people, and unembroiled in any controversy about the Yates' correction [for continuity]" [421, p. 910].

Feinstein put forth some advantages and disadvantages of permutation tests that were insightful for the time and foreshadowed later research. In terms of permutation tests, he listed five advantages:

1. The result of a permutation test is a direct, exact probability value for the random likelihood of the observed difference.
2. Permutation tests do not require any unwarranted inferential estimations of means, variances, pooled variances, or other parameters of an unobserved, hypothetical parent population. The tests are based solely on the evidence that was actually obtained.[42]
3. The investigator is not forced into making any erroneous assumptions either that the contrasted groups were chosen as random samples from a parent population, or that treatments under study were randomly allocated to the two groups.
4. The investigator is not forced into making any erroneous or unconfirmable assumptions about a Gaussian (or any other) distribution for the parent population, or about equal variances in the contrasted groups.
5. A permutation test can be applied to groups of any size, no matter how large or small. There are no degrees of freedom to be considered. In the case of a contingency table, there is no need to worry about the magnitude of the expected value, no need to calculate expectations based on fractions of people, and no need to worry about applying, or not applying, Yates correction for continuity.

[41] Here, Feinstein utilized permutation tests as the gold standard against which to evaluate classical tests, referencing a 1963 article by McHugh [914] and 1966 articles by Baker and Collier [52], and Edgington [388].

[42] In this second advantage, Feinstein clearly described the data-dependent nature of permutation tests, anticipating by many years later research on permutation methods.

Feinstein observed that while there were definite advantages to permutation tests, there were also disadvantages. The first three (of four) he considered as features that contributed to "the existing state of statistical desuetude" and labeled them inertia, ideology, and information [421, p. 911]:

1. Inertia: It is easier for many teachers to continue the inertia of teaching what they were taught years ago than to revise the contents of their lectures.
2. Ideology: Investigators who ideologically believe that the goal of science is to estimate parameters and variances will have no enthusiasm for tests that do not include or rely on these estimations.
3. Information: Many investigators have a deep-seated horror of doing anything that might entail losing information.
4. Permutation tests are notoriously difficult to calculate.

Feinstein elaborated on Items 3 and 4. Regarding Item 3, he emphasized that a loss of information would occur if raw data were converted into ordinal ranks for the sake of a non-parametric test that analyzes ranks rather than the observed raw scores. He explained that since ranks are used in nearly all non-parametric tests and since all non-parametric tests depend on random permutations, a statistician may erroneously conclude that all non-parametric tests create a loss of information.[43] He retorted that that conclusion was specious as "the non-parametric permutation tests illustrated here make use of the original values of the [observed] data, not the ranks" [421, p. 911].

Regarding Item 4, Feinstein observed that every permutation test must be computed entirely from the individual values of the observed data. Thus, each application is a unique test and precludes the compilation of tables that can be used repeatedly [421, p. 912]; a point made earlier, and most emphatically, by Bradley [201]. He followed this with the prescient observation that "in the era of the digital computer … these calculational difficulties will ultimately disappear" [421, p. 912]. Feinstein further observed that in situations where the sample sizes were large, the exact permutation test could be "truncated" into a Monte Carlo (resampling) type of test.

In a strongly worded conclusion, Feinstein argued that the ultimate value of permutation tests was that their intellectual directness, precision, and simplicity would free both the investigator and the statistician from "a deleterious pre-occupation with sampling distributions, pooled variances, and other mathematical distractions" [421, p. 914]. Finally, he noted that "an investigator who comprehends the principles of his statistical tests will be less inclined to give idolatrous worship to a numerical 'significance' that has no scientific connotation" [421, p. 914].[44]

[43] In the literature of mathematical statistics there are examples of distributions where a non-parametric test that "throws away information" is clearly superior to a parametric test; see for example, articles by Festinger in 1946 [427], Pitman in 1948 [1132], Whitney in 1948 [1445], and van den Brink and van den Brink in 1989 [1389].

[44] See also an informative and engaging 2012 article on this topic by Megan Higgs in *American Scientist* [616].

4.14 The Mann–Whitney, Pitman, and Cochran Tests

In 1973 Dinneen and Blakesley published algorithm and FORTRAN subroutine UDIST for generating the sampling distribution of the Mann–Whitney U statistic [351]. The algorithm differed from previous algorithms in that it produced a complete distribution of all possible U statistics instead of just a probability value. In addition, the algorithm proved to be twenty times faster than a previous algorithm published by Odeh in 1972 [1048]. Exact results were provided for small samples ($m, n \leq 10$) and accurate results to nine decimal places for larger sample sizes ($m, n \leq 79$), where m and n denote the two sample sizes.

In 1975 Arbuckle and Aiken published a FORTRAN program for Pitman's two-sample test for differences in location [30]. Output from the program consisted of the value of the differences between the means of the two samples, $\bar{x}$ and $\bar{y}$, and one- and two-tailed exact probability values. Such an approach was also supported by Odén and Wedel in 1975, who argued for the use of the two-sample permutation test over the conventional Student two-sample t test [1049].

In the same year, Patil presented "a relatively simple method for computing an exact null and nonnull distribution of [Cochran's] Q [test]" [1090, p. 186]. Reflecting the practice of defining permutation tests as the gold standard, Patil noted that it was now "possible to assess the performance of the asymptotic distribution" [1090, p. 189]. Also in 1975 Radlow and Alf proposed an exact chi-squared test in response to an article by Tate and Hyer 2 years previously [1341], where they noted that Tate and Hyer had compared results from chi-squared goodness-of-fit tests with results from exact multinomial tests, but had ordered terms by their probability values instead of by their discrepancies from the null hypothesis [1150]. Radlow and Alf concluded that an exact chi-squared test should be used whenever expected cell frequencies were small.

4.15 Mielke–Berry–Johnson and MRPP

In 1976 Mielke, Berry, and Johnson introduced a class of data-dependent tests termed multi-response permutation procedures (MRPP) based on distance functions [971]. As will be noted (q.v. page 254), MRPP are able to avoid the robustness problems associated with the use of squared Euclidean distances associated with univariate and multivariate analysis of variance methods by using ordinary Euclidean distances.

P.W. Mielke

Paul W. Mielke Jr. received his B.A. degree in mathematics from the University of Minnesota in 1953. In 1953–1954 he was trained in meteorology at the University of Chicago for the United States Air Force. After completing

(continued)

his military tour of duty in 1957, he resumed his academic career earning his M.A. degree in mathematics from the University of Arizona in 1958 and his Ph.D. in biostatistics from the University of Minnesota in 1963. Mielke accepted an appointment in the Department of Mathematics and Statistics at Colorado State University in 1963, where he remained until his retirement in 2002.

K.J. Berry

Kenneth J. Berry received his B.A. degree in sociology from Kalamazoo College in 1962 and his Ph.D. in sociology from the University of Oregon in 1966. He was employed by the State University of New York at Buffalo (now, University of Buffalo) from 1966 to 1970 and then joined the Department of Sociology at Colorado State University, where he remained for the rest of his academic career.

E.S. Johnson

Earl S. Johnson received his Ph.D. in statistics from Colorado State University in 1973 and worked for most of his career at Norwich Pharmaceuticals.

Based on ordinary Euclidean distances rather than the usual squared Euclidean distances, MRPP provided highly-robust, distribution-free, multivariate, Euclidean-distance-based permutation alternatives to analyzing experimental designs that normally employed classical analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA) analyses [940, 978]. The Mielke et al. 1976 article was motivated by a study sponsored by the National Communicable Disease Center[45] and involved comparisons of proportional contributions of five plague organism protein bands based on electrophoresis measurements obtained from samples of organisms associated with distinct geographical locations, such as Colorado, Vietnam, and Tasmania. Thus this initial example motivating MRPP involved comparisons of five-dimensional multivariate data sets. As P.K. Sen noted, "one of the beauties of . . . permutation tests is their distribution freeness even for the mulivariate [sic] distributions (where the unconditional tests usually fail to do so)" [1246, p. 1210]. MRPP were capable of incorporating both squared deviations from the mean and absolute deviations from the median.

[45] From 1946 to 1967, it was called the Communicable Disease Center (CDC), in 1967 it was renamed the National Communicable Disease Center, in 1970 it was again renamed the Center for Disease Control, in 1980 it became the Centers for Disease Control, and finally in 1992, the Centers for Disease Control and Prevention.

R.J. Boscovich and Absolute Deviations

The analysis of observations using absolute deviations or differences has a long and distinguished history in the statistical literature; see for example, articles by deLaubenfels [340], Eisenhart [407], Farebrother [419], Sheynin, [1264], and Stigler [1317]. Its beginnings in the eighth century were in the general area of regression analysis. As described by Sheynin, the earliest known use of regression was by Daniel Bernoulli (circa 1734) for astronomical prediction problems that involved the use of least (sum of) absolute deviations (LAD) regression [1264, p. 310]. In 1757 the Croatian Jesuit Roger Joseph (Rogerius Josephus) Boscovich, one of the last polymaths and an astronomer, formulated the principle that, given paired values of variables x and y connected by a linear relationship of the form $y = \alpha + \beta x$, the values of a and b that should be adopted for α and β, respectively, so that the line $y = a + bx$ that is most nearly in accord with the observations should be determined by two conditions:

1. The sums of the positive and negative corrections (to the y-values) shall be equal.
2. The sum of the (absolute values of) all of the corrections, positive and negative, shall be as small as possible.

(Boscovich, quoted in Eisenhart [407, p. 200]; see also discussions by Sheynin in 1973 [1264, p. 307] and Farebrother in 2001 [420]). This was a bold move that defied the canons of statistical methods of the time that relied on squared deviations from the regression line and challenged the conventional wisdom of least-squares analysis.

As noted by George Barnard in a series of lectures at Eidgenössische Technische Hochschule (ETH) Zürich during the winter of 1982, the origin of a least absolute estimator can be traced back to Galileo Galilei in 1632 [354, 1192]. In his "Dialogo due massimi sistemi" [492], Galilei considered the question of determining the distance from the earth of a new star, given observations on its maximum and minimum elevation and the elevation of the pole star by 13 observers at different points on the earth's surface [589, p. 148]. Galilei proposed the least possible correction in order to obtain a reliable result for the problem.

In 1789 Pierre-Simon Laplace gave an algebraic formulation and derivation of Boscovich's equation

$$\sum_{i=1}^{n} \big|(y_i - \bar{y}) - b(x_i - \bar{x})\big| = \text{minimum} ,$$

where $\bar{x}$ and $\bar{y}$ are arithmetic means, commenting that "Boscovich has given for this purpose an ingenious method" (Laplace, quoted in Eisenhart [407, p. 204]). Laplace clearly recognized the value of Boscovich's procedure and

(continued)

employed it on several occasions. Laplace compared the method of least-squares, which he called "the most advantageous method," with the method of Boscovich, which he called "the method of situation," concluding that the method of least-squares produced better predictions when the errors were normally-distributed. Nathaniel Bowditch, an American mathematician, navigator, and astronomer, in his translation of Laplace's *Mécanique Céleste*, noted in a footnote that the "method, proposed by Boscovich . . . is not now so much used as it ought to be" and in that same footnote, he added

> We shall hereafter find . . . , that the method of least-squares, when applied to a system of observations, in which one of the extreme errors is very great, does not generally give so correct a result as the method proposed by Boscovich . . . . The reason is, that in the former method, this extreme error affects the result in proportion to the *second* power of the error; but in the other method, it is as the *first* power, and must therefore be less [802, p. 438] (Bowditch, quoted in Eisenhart [407, p. 208] and also in Sheynin [1264, p. 311]).

In 1887 and again in 1923 Francis Ysidro Edgeworth dropped Condition 1 of Boscovich—that the sums of the positive and negative deviations be equal—and used Condition 2—that the sum of the absolute values of the deviations be a minimum. Thus, Edgeworth recommended the use of the median instead of the arithmetic mean so as to reduce the influence of "discordant" observations [381–386]; see also A.L. Bowley's tribute to Edgeworth in 1928 [189].

Edgeworth, examining Condition 2 of Boscovich, devised what he called a "double median" method for determining values of a and b that corresponded to the minimum of the sum

$$\sum_{i=1}^{n} |y_i - a - bx_i| \ ,$$

where the median of an odd number of observations $y_1, \ y_2, \ldots, y_n$ was the solution to

$$\sum_{i=1}^{n} |y_i - a| = \text{minimum}$$

[407, p. 208]; see also an informative 1997 article by Portnoy and Koenker [1142, p. 281]. As Eisenhart summarized, if the purpose is to minimize the apparent inconsistency of a set of observations as measured by some simple function of their residuals, then practical requirements of objectivity, applicability, unique solutions, and computational simplicity lead to adoption of the principle of least (sum of) squared residuals, and it was for these reasons, Eisenhart surmised, that the method of least-squares rapidly

(continued)

pushed Boscovich's method into the background [407, p. 209]. Roger Joseph Boscovich F.R.S passed away on 13 February 1787 in Milan, Italy.

The ascendency of the method of least-squares, to the almost complete exclusion of all other procedures, was greatly aided by the independent formulation, development, and publication of the method by the French mathematician Adrien Marie Legendre in 1805 and Carl Friedrich Gauss in 1809. Legendre, while not the first to use the method of least-squares, was the first to publish it [589, p. 152]. Legendre, in an appendix "On the method of least squares" in his 1805 book titled *Nouvelles méthodes pour la détermination des orbites des comètes* (*New Methods for Determining the Orbits of Comets*) introduced the technique of least (sum of) squared residuals and deduced the rules for forming the normal equations. Gauss, on the other hand, claimed priority in the use of the method of least-squares [589, pp. 153–154].[46] For more recent work on minimizing the sum of absolute deviations, see papers by Rhodes in 1930 [1165], Singleton in 1940 [1279], and Harris in 1950 [588].

Permutation methods, which are intrinsically distribution-free, are ideally suited to the use of absolute deviations and differences, replacing conventional squared deviations and differences. As noted by Westgard and Hunt in 1973, the results of an analysis based on least squares can be invalidated by one or two errant data points, and the least-squares results may also be inaccurate when the random error is large and the range of the data is small [1438, p. 53]. On this topic, see also articles by Harter in 1974 [589, p. 168], Hampel, Ronchetti, Rousseeuw, and Stahel in 1986 [582, p. 309], and Ronchetti in 1987 [1192, pp. 67–68]. Because permutation methods are data-dependent and do not require assumptions such as normality, with the attendant requirement of squared deviations from the mean, the choice of how to measure differences among observations is unrestricted. However, a Euclidean distance, being a metric, is most defensible. The advantage of using Euclidean distances among observations is that they minimize the impact of extreme observations, thereby creating a robust alternative to squared Euclidean distances among observations.[47] Moreover, squared Euclidean distances among observations yield a non-metric analysis space, whereas Euclidean distances among observations yield a metric analysis space that is congruent with the data space usually in question. The use of Euclidean distances among observations has proliferated in recent decades, with applications in linear regression and comparisons of treatments and groups, both univariate and multivariate.

[46] Stigler provides an excellent discussion of the priority of least-squares analysis [1321, pp. 320–331]; see also Gigerenzer, Swijtink, Porter, and Daston [512, pp. 80–84], as well as Maltz [873].

[47] Tukey refers to the inclusion of extreme observations in distributions of measurements as "contaminated distributions" [1379].

4.15.1 Least Absolute Deviations Regression

One application involving Euclidean distances among observations deserves special mention: least absolute deviations (LAD) regression analysis. While ordinary least-squares (OLS) regression has long been a staple for many research applications, the optimal properties of estimators of OLS regression are achieved only when the errors are normally-distributed. LAD regression is an attractive alternative when the errors are not normally-distributed; see for example, papers by Blattberg and Sargent in 1971 [171]; Gentle, and Narula and Wellington in 1977 [506,1022,1023]; Bassett and Koenker, Pfaffenberger and Dinkel, and Wilson in 1978 [84,1123,1462]; Narula and Wellington in 1979 [1024]; Bloomfield and Steiger in 1980 [173]; Wellington and Narula in 1981 [1434]; Koenker and Bassett, and Narula and Wellington in 1982 [765, 1025]; Seneta in 1983 [1248]; Seneta and Steiger in 1984 [1249]; Dielman in 1986 [349]; Dielman in 1989 [350]; Hurvich and Tsai in 1990 [670]; Mathew and Nordström in 1993 [904]; and Cade and Richards in 1996 [233]. In addition, LAD regression is much less sensitive to the inclusion of extreme values as the errors are not squared [152].[48]

Consider a simple linear regression model with a single predictor variable (x) and a single criterion variable (y) with n paired x_i and y_i observed values for $i = 1, \ldots, n$. The LAD regression equation is given by

$$\tilde{y}_i = \tilde{\alpha} + \tilde{\beta} x_i \ ,$$

where $\tilde{y}_i$ is the ith of n predicted values, x_i is the ith of n predictor values, and $\tilde{\alpha}$ and $\tilde{\beta}$ are the least absolute parameter estimates of the intercept and slope, respectively. Estimates of LAD regression parameters are computed by minimizing the sum of the absolute differences between the observed y_i and predicted $\tilde{y}_i$ values for $i = 1, \ldots, n$; viz.,

$$\sum_{i=1}^{n} \left| y_i - \tilde{y}_i \right| \ .$$

Unlike OLS regression, no closed-form expressions can be given for $\tilde{\alpha}$ and $\tilde{\beta}$; however, values for $\tilde{\alpha}$ and $\tilde{\beta}$ may be obtained via linear programming, as detailed by Barrodale and Roberts in 1973 and 1974 [74, 75].

4.15.2 Multi-Response Permutation Procedures

Let $\Omega = \{\omega_1, \ldots, \omega_N\}$ be a finite sample of N objects that is representative of some target population in question. Let $x_I' = [x_{1I}, \ldots, x_{rI}]$ be a transposed vector

[48] The 1977(4) issue of *Communications in Statistics—Simulation and Computation*, edited by James E. Gentle, was devoted to computations for least absolute values estimation and is an excellent source for an introduction to LAD regression [506].

of r commensurate response measurements for object ω_I, $I = 1, \ldots, N$, and let $S_1, \ldots, S_{g+1}$ designate an exhaustive partitioning of the N objects comprising Ω into $g + 1$ disjoint groups. Note that the response measurements can consist of either rank-order statistics or interval measurements, or any combination of the two. Also, let $\Delta_{I,J}$ be a symmetric distance function value of the response measurements associated with objects ω_I and ω_J, i.e.,

$$\Delta_{I,J} = \left[\sum_{h=1}^{r} \Big| x_{hI} - x_{hJ} \Big|^{p} \right]^{v/p} , \tag{4.2}$$

where x_{hI} and x_{hJ} are the hth coordinates of observations I and J in an r-dimensional space. The Minkowski family of metrics occurs when $v = 1$ and $p \geq 1$ [997]. If $v > 0$, $r \geq 2$, and $p = 2$, then $\Delta_{I,J}$ is rotationally invariant. When $v = 1$ and $p = 1$, $\Delta_{I,J}$ is a city-block metric, which is not rotationally invariant. When $v = 1$ and $p = 2$, $\Delta_{I,J}$ is the metric known as Euclidean distance. If $v = 2$ and $p = 2$, then $\Delta_{I,J}$ is a squared Euclidean distance, which is not a metric since the triangle inequality is not satisfied.[49] In this 1976 article, $p = 2$ and $v = 1$, yielding a Euclidean distance. The present form of the MRPP statistic, introduced by O'Reilly and Mielke in 1980 [1070], is given by

$$\delta = \sum_{i=1}^{g} C_i \xi_i \ , \tag{4.3}$$

where $C_i > 0$ is a classified group weight for $i = 1, \ldots, g$, $\sum_{i=1}^{g} C_i = 1$,

$$\xi_i = \binom{n_i}{2}^{-1} \sum_{I<J} \Delta_{I,J} \, \Psi_i(\omega_I) \, \Psi_i(\omega_J) \tag{4.4}$$

is the average distance function value for all distinct pairs of objects in group S_i for $i = 1, \ldots, g$, $n_i \geq 2$ is the number of a priori objects classified into group S_i for $i = 1, \ldots, g$, $K = \sum_{i=1}^{g} n_i$, $n_{g+1} = N - K \geq 0$ is the number of remaining (unclassified) objects in an "excess" group S_{g+1} that is an empty group in most applications, $g \geq 1$ where $N > k$ if $g = 1$, $\sum_{I<J}$ is the sum over all I and J such that $1 \leq I < J \leq N$, and $\Psi(\cdot)$ is an indicator function given by

$$\Psi_i(\omega_I) = \begin{cases} 1 & \text{if } \omega_I \in S_i \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

[49] A distance function is a metric if it satisfies three properties given by (1) $\Delta_{I,J} \geq 0$ and $\Delta_{I,I} = 0$, (2) $\Delta_{I,J} = \Delta_{J,I}$ (i.e., symmetry), and (3) $\Delta_{I,J} \leq \Delta_{I,K} + \Delta_{K,J}$ (i.e., the triangle inequality).

Incidentally, $\Delta_{I,J} = \min(\Delta_{I,J}, B)$ where $B > 0$, a specified truncation constant, has been found useful for detecting events such as multiple clumping of response measurements within groups [943, 965, pp. 40–44].

The choice of the classified group weights, $C_1, \ldots, C_g$, and the symmetric distance function, $\Delta_{I,J}$, specify the structure of MRPP. While Mielke, Berry, and Johnson [971] restricted C_i to

$$C_i = \frac{n_i(n_i - 1)}{\displaystyle\sum_{j=1}^{g} n_j(n_j - 1)}$$

for $i = 1, \ldots, g$, other group weights could be considered. For example

$$C_i = \frac{n_i}{K}\,, \quad C_i = \frac{n_i - 1}{K - g}\,, \quad \text{and} \quad C_i = \frac{1}{g}$$

for $i = 1, \ldots, g$. In 1970, a paper by Mantel and Valand introduced an early version of MRPP [889]. There were three problems with the approach of Mantel and Valand. First, Mantel and Valand used a city-block distance function which was not invariant to coordinate rotation. Second, they chose the same inefficient group weight as Mielke, Berry, and Johnson [971], i.e.,

$$C_i = \frac{n_i(n_i - 1)}{\displaystyle\sum_{j=1}^{g} n_j(n_j - 1)}\,,$$

where C_i is the group weight for the ith of g groups and n_i is the number of objects in the ith group. Third, Mantel and Valand erroneously used a U-statistic argument based on a paper by Hoeffding published in 1948 [637] that claimed that the distribution of the statistic was asymptotically normal [935, 936].

Although not part of the 1976 article by Mielke, Berry, and Johnson, it should be noted that when $r = 1$, $v = p = 2$, $K = N$, and $C_i = (n_i - 1)/(N - g)$ for $i = 1, \ldots, g$, δ is a permutation version of the squared two-sample t and the one-way analysis of variance F statistics, commonly termed Fisher–Pitman permutation tests [451, 1131]. Here, the identity specifying the association between F and δ is given by

$$\delta = \frac{2\left[N\displaystyle\sum_{I=1}^{N} x_I^2 - \left(\sum_{I=1}^{N} x_I\right)^2\right]}{N[N - g + (g + 1)F]}$$

and x_I is the response measurement for the Ith of N objects. Alternatively,

$$\delta = \frac{2MS_{\text{Between}}}{F} \qquad \text{and} \qquad F = \frac{2MS_{\text{Between}}}{\delta} \ ,$$

where

$$MS_{\text{Between}} = \frac{1}{g-1} \sum_{i=1}^{g} \left(\bar{x}_i - \bar{\bar{x}} \right)^2 \ ,$$

$\bar{x}_i$ is the mean of the ith of g groups, and $\bar{\bar{x}}$ is the grand mean of all objects, i.e.,

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \qquad \text{and} \qquad \bar{\bar{x}} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} x_{ij} \ ,$$

where x_{ij} is the response measurement on the ith object in the jth group in the usual alternate univariate analysis of variance notation; see also a 1982 paper by Mielke, Berry, and Medina [981, p. 790] and a discussion in a 2007 book by Mielke and Berry [965, p. 51]. When interval data are replaced with rank-order statistics, then δ includes rank tests such as the Kruskal–Wallis analysis of variance rank test test. When interval data are replaced with rank-order statistics and $g = 2$, then δ includes the Wilcoxon [1453], Festinger [427], Mann–Whitney [880], Whitfield [1443], Haldane–Smith [573], and van der Reyden [1391] two-sample rank-sum tests; see also on this topic, a 1981 paper by Mielke, Berry, Brockwell, and Williams [969]. The robustness of $v = 1$ over $v = 2$ is demonstrated via examples in Sect. 6.16 in Chap. 6.

The null hypothesis states that equal probabilities are assigned to each of the

$$M = \frac{N!}{\displaystyle\prod_{i=1}^{g+1} n_i!}$$

possible allocations of the N objects in Ω to the $g + 1$ groups, $S_1, \ldots, S_{g+1}$. Under the null hypothesis, the N multi-response measurements are exchangeable multivariate random variables (q.v. page 4). The probability associated with an observed value of δ, say δ_{o}, is the probability under the null hypothesis of observing a value of δ as extreme or more extreme than δ_{o}. Thus, an exact probability value for δ_{o} may be expressed as

$$P(\delta \leq \delta_{\text{o}} | H_0) = \frac{\text{number of } \delta \text{ values} \leq \delta_{\text{o}}}{M}$$

or

$$P(\delta \geq \delta_{\text{o}} | H_0) = \frac{\text{number of } \delta \text{ values} \geq \delta_{\text{o}}}{M} \ .$$

Table 4.4 Example data set with $g = 2$, $r = 2$, $N = 7$, $n_1 = 4$, and $n_2 = 3$

		Values				Values	
Group	Object	x_1	x_2	Group	Object	x_1	x_2
S_1	ω_1	4	1	S_2	ω_5	2	2
S_1	ω_2	5	6	S_2	ω_6	2	3
S_1	ω_3	5	4	S_2	ω_7	3	2
S_1	ω_4	4	3				

4.15.3 An Example MRPP Analysis

To illustrate the computation of MRPP, consider a finite sample of $N = 7$ objects and let S_1 and S_2 denote an exhaustive partitioning of the N objects into $g = 2$ groups. Further, let S_1 consist of $n_1 = 4$ objects with $r = 2$ measurements (x_{1I} and x_{2I}) on each object for $I = 1, \ldots, 7$, with $x_1' = \{4, 1\}$, $x_2' = \{5, 6\}$, $x_3' = \{5, 4\}$, and $x_4' = \{4, 3\}$, and let S_2 consist of $n_2 = 3$ objects with $r = 2$ measurements on each object, with $x_5' = \{2, 2\}$, $x_6' = \{2, 3\}$, and $x_7' = \{3, 2\}$. Here the numbers are deliberately made small to facilitate the example analysis. The multivariate data for the $N = 7$ objects are listed in Table 4.4. For this example, let $v = 1$, $p = 2$, $C_1 = n_1/N = 4/7$, and $C_2 = n_2/N = 3/7$, so that the two groups are weighted proportional to their sizes, and $K = N$. Then following Eq. (4.2) for group S_1 with $n_1 = 4$ objects,

$$\Delta_{1,2} = \left[(4-5)^2 + (1-6)^2\right]^{1/2} = 5.0990 \,,$$

$$\Delta_{1,3} = \left[(4-5)^2 + (1-4)^2\right]^{1/2} = 3.1623 \,,$$

$$\Delta_{1,4} = \left[(4-4)^2 + (1-3)^2\right]^{1/2} = 2.0000 \,,$$

$$\Delta_{2,3} = \left[(5-5)^2 + (6-4)^2\right]^{1/2} = 2.0000 \,,$$

$$\Delta_{2,4} = \left[(5-4)^2 + (6-3)^2\right]^{1/2} = 3.1623 \,,$$

and

$$\Delta_{3,4} = \left[(5-4)^2 + (4-3)^2\right]^{1/2} = 1.4142 \,.$$

For group S_2 with $n = 3$ objects,

$$\Delta_{5,6} = \left[(2-2)^2 + (2-3)^2\right]^{1/2} = 1.0000\ ,$$

$$\Delta_{5,7} = \left[(2-3)^2 + (2-2)^2\right]^{1/2} = 1.0000\ ,$$

and

$$\Delta_{6,7} = \left[(2-3)^2 + (3-2)^2\right]^{1/2} = 1.4142\ .$$

Then following Eq. (4.4),

$$\begin{aligned}
\xi_1 &= \binom{n_1}{2}^{-1} (\Delta_{1,2} + \Delta_{1,3} + \Delta_{1,4} + \Delta_{2,3} + \Delta_{2,4} + \Delta_{3,4}) \\
&= \binom{4}{2}^{-1} (5.0990 + 3.1623 + 2.0000 + 2.0000 + 3.1623 + 1.4142) \\
&= 2.8063\ , \\
\xi_2 &= \binom{n_2}{2}^{-1} (\Delta_{5,6} + \Delta_{5,7} + \Delta_{6,7}) \\
&= \binom{3}{2}^{-1} (1.0000 + 1.0000 + 1.4142) \\
&= 1.1381\ ,
\end{aligned}$$

and the weighted mean as defined in Eq. (4.3) is

$$\delta = C_1\xi_1 + C_2\xi_2 = \left(\frac{4}{7}\right)(2.8063) + \left(\frac{3}{7}\right)(1.1381) = 2.0903\ .$$

Smaller values of δ indicate a concentration of response measurements within the g groups, whereas larger values of δ indicate a lack of concentration between response measurements among the g groups [968]. The $N = 7$ objects can be partitioned into $g = 2$ groups, S_1 and S_2, with $n_1 = 4$ and $n_2 = 3$, respectively, in precisely

$$M = \frac{N!}{n_1!\, n_2!} = \frac{7!}{4!\ 3!} = 35$$

ways. The 35 permutations obtained from the observed data set in Table 4.4 along with ξ_1, ξ_2, and δ values are listed in Table 4.5 and are ordered from lowest to highest

Table 4.5 Permutations of the observed data set in Table 4.4 for groups S_1 and S_2 with values for ξ_1, ξ_2, and δ, ordered by values of δ

Order	S_1	S_2	ξ_1	ξ_2	δ
1	$\{(4,1)\,(2,2)\,(2,3)\,(3,2)\}$	$\{(5,6)\,(5,4)\,(4,3)\}$	1.6488	2.1922	1.8817
2	$\{(4,1)\,(5,6)\,(5,4)\,(4,3)\}$	$\{(2,2)\,(2,3)\,(3,2)\}$	2.8063	1.1381	2.0913
3	$\{(5,6)\,(5,4)\,(4,3)\,(2,3)\}$	$\{(4,1)\,(2,2)\,(3,2)\}$	2.6636	1.5501	2.1864
4	$\{(5,6)\,(5,4)\,(4,3)\,(3,2)\}$	$\{(4,1)\,(2,2)\,(2,3)\}$	2.5485	2.0215	2.3227
5	$\{(4,1)\,(4,3)\,(2,2)\,(3,2)\}$	$\{(5,6)\,(5,4)\,(2,3)\}$	1.7168	3.1350	2.3246
6	$\{(4,3)\,(2,2)\,(2,3)\,(3,2)\}$	$\{(4,1)\,(5,6)\,(5,4)\}$	1.5107	3.4204	2.3292
7	$\{(5,6)\,(5,4)\,(4,3)\,(2,2)\}$	$\{(4,1)\,(2,3)\,(3,2)\}$	2.9030	1.8856	2.4670
8	$\{(4,1)\,(4,3)\,(2,2)\,(2,3)\}$	$\{(5,6)\,(5,4)\,(3,2)\}$	2.0501	3.1002	2.5001
9	$\{(5,6)\,(5,4)\,(2,2)\,(2,3)\}$	$\{(4,1)\,(4,3)\,(3,2)\}$	3.1684	1.6095	2.5003
10	$\{(4,1)\,(5,6)\,(5,4)\,(3,2)\}$	$\{(4,3)\,(2,2)\,(2,3)\}$	3.1627	1.7454	2.5553
11	$\{(4,1)\,(4,3)\,(2,3)\,(3,2)\}$	$\{(5,6)\,(5,4)\,(2,2)\}$	1.8452	3.5352	2.5695
12	$\{(5,6)\,(2,2)\,(2,3)\,(3,2)\}$	$\{(4,1)\,(5,4)\,(4,3)\}$	2.8548	2.1922	2.5708
13	$\{(4,1)\,(5,6)\,(5,4)\,(2,3)\}$	$\{(4,3)\,(2,2)\,(3,2)\}$	3.4158	1.5501	2.6162
14	$\{(4,1)\,(5,4)\,(4,3)\,(3,2)\}$	$\{(5,6)\,(2,2)\,(2,3)\}$	2.0389	3.4142	2.6283
15	$\{(5,6)\,(5,4)\,(2,3)\,(3,2)\}$	$\{(4,1)\,(4,3)\,(2,2)\}$	3.0199	2.1574	2.6503
16	$\{(4,1)\,(5,6)\,(5,4)\,(2,2)\}$	$\{(4,3)\,(2,3)\,(3,2)\}$	3.5172	1.6095	2.6996
17	$\{(4,1)\,(5,4)\,(2,2)\,(3,2)\}$	$\{(5,6)\,(4,3)\,(2,3)\}$	2.3744	3.1350	2.7004
18	$\{(5,4)\,(2,2)\,(2,3)\,(3,2)\}$	$\{(4,1)\,(5,6)\,(4,3)\}$	2.1684	3.4204	2.7050
19	$\{(5,6)\,(4,3)\,(2,2)\,(2,3)\}$	$\{(4,1)\,(5,4)\,(3,2)\}$	2.9402	2.4683	2.7379
20	$\{(4,1)\,(5,6)\,(2,2)\,(2,3)\}$	$\{(5,4)\,(4,3)\,(3,2)\}$	3.4010	1.8856	2.7516
21	$\{(4,1)\,(5,6)\,(2,2)\,(3,2)\}$	$\{(3,4)\,(4,3)\,(2,3)\}$	3.2036	2.1922	2.7701
22	$\{(5,6)\,(5,4)\,(2,2)\,(3,2)\}$	$\{(4,1)\,(4,3)\,(2,3)\}$	3.1510	2.2761	2.7761
23	$\{(4,1)\,(5,6)\,(4,3)\,(3,2)\}$	$\{(5,4)\,(2,2)\,(2,3)\}$	2.9270	2.5893	2.7822
24	$\{(4,1)\,(5,4)\,(2,2)\,(2,3)\}$	$\{(5,6)\,(4,3)\,(3,2)\}$	2.6658	3.0162	2.8160
25	$\{(4,1)\,(5,4)\,(4,3)\,(2,2)\}$	$\{(5,6)\,(2,3)\,(3,2)\}$	2.4424	3.3763	2.8426
26	$\{(5,4)\,(4,3)\,(2,2)\,(2,3)\}$	$\{(4,1)\,(5,6)\,(3,2)\}$	2.2364	3.6618	2.8473
27	$\{(5,6)\,(4,3)\,(2,3)\,(3,2)\}$	$\{(4,1)\,(5,4)\,(2,2)\}$	2.7842	3.0013	2.8773
28	$\{(4,1)\,(5,4)\,(4,3)\,(2,3)\}$	$\{(5,6)\,(2,2)\,(3,2)\}$	2.4279	3.4907	2.8834
29	$\{(4,1)\,(5,6)\,(2,3)\,(3,2)\}$	$\{(3,4)\,(4,3)\,(2,2)\}$	3.2451	2.4186	2.8909
30	$\{(4,1)\,(5,4)\,(2,3)\,(3,2)\}$	$\{(5,6)\,(4,3)\,(2,2)\}$	2.4683	3.4661	2.8959
31	$\{(4,1)\,(5,6)\,(4,3)\,(2,3)\}$	$\{(5,4)\,(2,2)\,(3,2)\}$	3.2221	2.4780	2.9032
32	$\{(5,4)\,(4,3)\,(2,3)\,(3,2)\}$	$\{(4,1)\,(5,6)\,(2,2)\}$	2.0389	4.1117	2.9272
33	$\{(5,4)\,(4,3)\,(2,2)\,(3,2)\}$	$\{(4,1)\,(5,6)\,(2,3)\}$	2.0831	4.0567	2.9289
34	$\{(4,1)\,(5,6)\,(4,3)\,(2,2)\}$	$\{(5,4)\,(2,3)\,(3,2)\}$	3.2889	2.4683	2.9372
35	$\{(5,6)\,(4,3)\,(2,2)\,(3,2)\}$	$\{(4,1)\,(5,4)\,(2,3)\}$	2.8808	3.0510	2.9537

by the δ values. The observed statistic, $\delta_{\text{o}} = 2.0913$, obtained for the realized partition is unusual since 33 of the remaining δ values exceed the observed δ_{o} value of 2.0913 and only one value of δ is smaller: $\delta = 1.8817$. If all partitions occur with equal chance, the exact probability value of $\delta_{\text{o}} = 2.0903$ is

$$P(\delta \leq \delta_{\mathrm{o}}|H_0) = \frac{\text{number of } \delta \text{ values} \leq \delta_{\mathrm{o}}}{M} = \frac{2}{35} = 0.0571 \;.$$

4.15.4 Approximate Probability Values

The 1976 article by Mielke, Berry, and Johnson [971] provided a useful moment-approximation for the distribution of δ, standardizing δ by

$$T = \frac{\delta - \mu_\delta}{\sigma_\delta}$$

and approximating the distribution of

$$T_B = T\left[\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\right]^{1/2} + \frac{\alpha}{\alpha+\beta}$$

with the beta distribution having a density function given by

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\,\beta)} & \text{if } 0 < x < 1 \;, \\ 0 & \text{otherwise} \;, \end{cases}$$

where $\alpha > 0$, $\beta > 0$, and

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

Later, the distribution of T was approximated by the Pearson type III distribution [968, Sect. 5.2] since the Pearson type III distribution is completely characterized by the exact mean, variance, and skewness of δ under H_0 given by

$$\mu_\delta = \frac{1}{M}\sum_{I=1}^{M} \delta_I \;,$$

$$\sigma_\delta^2 = \frac{1}{M}\sum_{I=1}^{M} (\delta_I - \mu_\delta)^2 \;,$$

and

$$\gamma_\delta = \left[\frac{1}{M}\sum_{I=1}^{M} (\delta_I - \mu_\delta)^3\right] \Big/ \sigma_\delta^3 \;,$$

respectively. Efficient computational expressions for μ_δ, σ_δ^2, and γ_δ under H_0 were given by

$$
\begin{aligned}
\mu_\delta &= D(1)\ , \\
\sigma_\delta^2 &= 2\left\{\sum_{i=1}^{g} C_i^2\left[n_i^{(2)}\right]^{-1} - \left[N^{(2)}\right]^{-1}\right\}\left[D(2) - 2D(2') + D(2'')\right] \\
&\quad + 4\left[\sum_{i=1}^{g} C_i^2 n_i^{-1} - N^{-1}\right]\left[D(2') - D(2'')\right]\ , \\
\gamma_\delta &= \left\{E\left[\delta^3\right] - 3\mu_\delta\sigma_\delta^2 - \mu_\delta^3\right\}/\sigma_\delta^3\ \ ,
\end{aligned}
$$

and

$$
\begin{aligned}
E\left[\delta^3\right] &= 4\sum_{i=1}^{g} C_i^3\left[n_i^{(2)}\right]^{-2} D(3) \\
&\quad + 8\sum_{i=1}^{g} C_i^3 n_i^{(3)}\left[n_i^{(2)}\right]^{-3}\left[3D\left(3'\right) + D\left(3^*\right)\right] \\
&\quad + 8\sum_{i=1}^{g} C_i^3 n_i^{(4)}\left[n_i^{(2)}\right]^{-3}\left[3D\left(3^{**}\right) + D\left(3^{***}\right)\right] \\
&\quad + 6\sum_{i=1}^{g} C_i^2\left\{1 - C_i + C_i n_i^{(4)}\left[n_i^{(2)}\right]^{-2}\right\}\left[n_i^{(2)}\right]^{-1} D\left(3''\right) \\
&\quad + 12\sum_{i=1}^{g} C_i^2\left\{(1 - C_i)n_i^{(3)} + C_i n_i^{(5)}\left[n_i^{(2)}\right]^{-1}\right\}\left[n_i^{(2)}\right]^{-2} D\left(3'''\right) \\
&\quad + \sum_{i=1}^{g} C_i\left\{(1 - C_i)(1 - 2C_i) + 3C_i(1 - C_i)n_i^{(4)}\left[n_i^{(2)}\right]^{-2}\right. \\
&\quad + \left. C_i^2 n_i^{(6)}\left[n_i^{(2)}\right]^{-3}\right\} D\left(3''''\right),
\end{aligned}
$$

where

$$
N^{(c)} = \frac{N!}{(N-c)!}
$$

and the necessary model parameters are the twelve symmetric functions defined by

$$\begin{aligned}
D(1) &= \frac{1}{N^{(2)}} \sum \Delta_{J_1,J_2} \ , \\
D(2) &= \frac{1}{N^{(2)}} \sum \Delta_{J_1,J_2}^2 \ , \\
D(2') &= \frac{1}{N^{(3)}} \sum \Delta_{J_1,J_2} \Delta_{J_1,J_3} \ , \\
D(2'') &= \frac{1}{N^{(4)}} \sum \Delta_{J_1,J_2} \Delta_{J_3,J_4} \ , \\
D(3) &= \frac{1}{N^{(2)}} \sum \Delta_{J_1,J_2}^3 \ , \\
D(3') &= \frac{1}{N^{(3)}} \sum \Delta_{J_1,J_2}^2 \Delta_{J_1,J_3} \ , \\
D(3'') &= \frac{1}{N^{(4)}} \sum \Delta_{J_1,J_2}^2 \Delta_{J_3,J_4} \ , \\
D(3''') &= \frac{1}{N^{(5)}} \sum \Delta_{J_1,J_2} \Delta_{J_1,J_3} \Delta_{J_4,J_5} \ , \\
D(3'''') &= \frac{1}{N^{(6)}} \sum \Delta_{J_1,J_2} \Delta_{J_3,J_4} \Delta_{J_5,J_6} \ , \\
D(3^*) &= \frac{1}{N^{(3)}} \sum \Delta_{J_1,J_2} \Delta_{J_1,J_3} \Delta_{J_2,J_3} \ , \\
D(3^{**}) &= \frac{1}{N^{(4)}} \sum \Delta_{J_1,J_2} \Delta_{J_1,J_3} \Delta_{J_2,J_4} \ ,
\end{aligned}$$

and

$$D(3^{***}) = \frac{1}{N^{(4)}} \sum \Delta_{J_1,J_2} \Delta_{J_1,J_3} \Delta_{J_1,J_4} \ ,$$

where J_1, J_2, J_3, J_4, J_5, and J_6 denote distinct integers from 1 to N, and the sums are over all permutations of the indices. The primes and asterisks are used merely to designate the twelve distinct symmetric function model parameters. An additional comment pertains to the combinations of symmetric function model parameters given by $D(2')-D(2'')$ and $D(2)-2D(2')+D(2'')$ in σ_δ^2. Although $D(2')-D(2'')$ may be positive, negative, or zero, $D(2) - 2D(2') + D(2'')$ is nonnegative since

$$\begin{aligned}
&4\left[D(2) - 2D(2') + D(2'')\right] \\
&\qquad = \frac{1}{N^{(4)}} \sum \left(\Delta_{J_1,J_2} - \Delta_{J_1,J_3} - \Delta_{J_2,J_4} + \Delta_{J_3,J_4}\right)^2 \geq 0 \ .
\end{aligned}$$

If $C_i = n_i/K$ and $K = N$, then some efficiency is secured since the coefficient of $D(2') - D(2'')$ in σ_δ^2 is 0.

Because the calculation of the model parameters involves a constant multiple of N^6, alternative computation forms are needed for calculating the model parameters associated with μ_δ, σ_δ^2, and γ_δ. The following results provide efficient computation forms to obtain the twelve symmetric function model parameters [971]. If

$$d_{kJ} = \sum_{J'=1}^{N} \Delta_{J,J'}^{k}$$

and

$$d_k = \sum_{J=1}^{N} d_{kJ}$$

for $k = 1, 2$, and 3. Then,

$$D(1) = \frac{1}{N^{(2)}}\, d_1 \ ,$$

$$D(2) = \frac{1}{N^{(2)}}\, d_2 \ ,$$

$$D(3) = \frac{1}{N^{(2)}}\, d_3 \ ,$$

$$D(2') = \frac{1}{N^{(3)}} \left[\sum_{J=1}^{N} d_{1J}^2 - d_2 \right] \ ,$$

$$D(2'') = \frac{1}{N^{(4)}} \left[d_1^2 - 4N^{(3)} D(2') - 2d_2 \right] \ ,$$

$$D(3') = \frac{1}{N^{(3)}} \left[\sum_{J=1}^{N} d_{1J}\, d_{2J} - d_3 \right] \ ,$$

$$D(3'') = \frac{1}{N^{(4)}} \left[d_1 d_2 - 4N^{(3)} D(3') - 2d_3 \right] \ ,$$

$$D(3^*) = \frac{6}{N^{(3)}} \sum_{J_1<J_2<J_3} \Delta_{J_1,J_2}\, \Delta_{J_1,J_3}\, \Delta_{J_2,J_3} \ ,$$

$$D(3^{**}) = \frac{1}{N^{(4)}} \left[2 \sum_{J_1<J_2} \Delta_{J_1,J_2} d_{1J_1} d_{1J_2} - 2N^{(3)} D(3') - N^{(3)} D(3^*) - d_3 \right] \ ,$$

$$D(3^{***}) = \frac{1}{N^{(4)}} \left[\sum_{J=1}^{N} d_{1J}^{3} - 3N^{(3)} D(3') - d_3 \right] ,$$

$$D(3''') = \frac{1}{N^{(5)}} \left[N^{(3)} d_1 D(2') - 4N^{(4)} D(3^{**}) - 2N^{(4)} D(3^{***}) \right.$$
$$\left. - 4N^{(3)} D(3') - 2N^{(3)} D(3^{*}) \right] ,$$

and

$$D(3'''') = \frac{1}{N^{(6)}} \left[N^{(4)} d_1 D(2'') - 8N^{(5)} D(3''') - 4N^{(4)} D(3'') \right.$$
$$\left. - 8N^{(4)} D(3^{**}) \right] ,$$

where $\sum_{I<J<L}$ is the sum over all I, J, and L such that $1 \le I < J < L \le N$. Thus, the actual computations involve only a constant multiple of N^2 operations to obtain the model parameters associated with μ_δ and σ_δ^2 and a constant multiple of N^3 operations to obtain the model parameters associated with γ_δ; see also a 1993 article by Charles Davis [330].

Near the end of the decade in 1978 and 1979, Mielke published two more papers on MRPP [935, 936]. In the 1978 paper, Mielke addressed a conclusion in the 1970 paper by Mantel and Valand [889] that the asymptotic distribution of the MRPP δ statistic was normal, based on the U statistics discussed by Hoeffding in 1948 [637]. The approach in the 1978 article simply established that the asymptotic skewness of the δ statistic was a substantial negative value, thereby demonstrating that the distribution of δ was not normal [935]. The 1979 paper by Mielke further demonstrated that the MRPP statistics were not asymptotically normal for cases with unequal sample sizes [936] and examined in a 1982 paper by Brockwell, Mielke, and Robinson [220]. For more on MRPP in completely randomized designs, see a 1983 article on permutation tests by Robinson [1179]. Due to the increasing speed of computers during this period, resampling-approximation MRPP probability-value programs were developed in addition to the previously-described exact and moment-approximation MRPP probability-value programs [965].

Two general asymptotic distributional properties of δ are important to mention. First, asymptotic non-normality of δ occurs when N/K and NC_i/n_i for $i = 1, \ldots, g \ge 2$ converge to 1 as $N \to \infty$ [220]. Second, a necessary condition for the asymptotic normality of δ is that $D(2') - D(2'') > 0$ as $N \to \infty$ [1070]. Even when $N^{1/2}(\delta - \mu_\delta)$ is asymptotically non-degenerate normal, the convergence is often so slow that a normal approximation is grossly inadequate even when N is quite large. Moreover, there are situations where $N^{1/2}(\delta - \mu_\delta)$ is asymptotically degenerate and the non-degenerate limiting distribution of $N(\delta - \mu_\delta)$ is non-normal [220, 1070]. In this regard, see also a 1988 article by Denker and Puri published in *Advances in Applied Mathematics* [342].

Fig. 4.8 Notation for an observed 2 × 2 contingency table

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d$

4.16 Determining the Number of Contingency Tables

When executing an exact permutation test on a 2 × 2 contingency table such as a Fisher exact probability test, it is possible to determine the exact number of possible arrangements of cell frequencies, given the marginal frequency totals, prior to enumeration of all possible 2 × 2 contingency tables. Consider a 2 × 2 contingency table with cell frequencies $\{a, b, c, d\}$, row marginal frequency totals $\{a + b, c + d\}$, column marginal frequency totals $\{a + c, b + d\}$, and frequency total $\{a + b + c + d\}$, such as is illustrated in Fig. 4.8. Then the number of possible arrangements of cell frequencies, given the marginal frequency totals, is given by

$$M = \min(a + b\,, a + c) - \max(0, a - d) + 1\ .$$

In 1977 Gail and Mantel published a brief paper describing exact and approximate methods for determining the number of arrangements of cell frequencies (M) consistent with marginal frequency totals in $r \times c$ contingency tables [490]. The method proved invaluable to permutation researchers as it is often necessary to determine if an exact permutation test is feasible or a resampling approach will be required prior to conducting a permutation test. The exact method of Gail and Mantel was restricted to smaller tables such as 2 × 2, 2 × k, and 3 × 3 contingency tables, but the approximate method utilized a normal approximation that yielded an estimate of the number of arrangements of cell frequencies in $r \times c$ contingency tables, given fixed marginal frequency totals.

In the same year, Klotz and Teng noted that it was difficult to determine the number of arrangements of cell frequencies in multi-way contingency tables with fixed marginal frequency totals due to the lack of an easily computed, closed-form expression that related M to the size of the table and the marginal frequency totals. They devised a geometric approach based on paths in an r-dimensional space, utilizing planes and lattice points that made the determination of M feasible [760].

4.17 Soms and the Fisher Exact Permutation Test

In 1977 Andrew P. Soms published an algorithm for the discrete Fisher–Pitman permutation test for differences between two independent samples [1296]. As Soms described the algorithm, let $x_i, i = 1, \ldots, k_1$, and $y_i, i = 1, \ldots, k_2$, denote the observed values in random samples from populations 1 and 2, respectively, and let $\bar{x}$ and $\bar{y}$ indicate the sample means. It is desired to test, at level α, the hypothesis that populations 1 and 2 are identical against the alternative hypothesis that population 2 is stochastically smaller than population 1. Then, denote by N the number of samples of size k_2 that can be drawn from the combined set $\{x_1, \ldots, x_{k_1}, y_1, \ldots, y_{k_2}\}$, without replacement, with sample means less than or equal to $\bar{y}$ counted, and the null hypothesis rejected if

$$N \bigg/ \binom{k_1 + k_2}{k_2} \leq \alpha \ .$$

Soms noted that the permutation version of the Wilcoxon two-sample rank-sum test is carried out in exactly the same way, except that the sum of the average ranks is used in place of $\bar{y}$.

Soms provided a FORTRAN computer program that carried out the Fisher–Pitman and Wilcoxon two-sample rank-sum permutation tests for up to 10 distinct data values. As Soms described the limitations of the program, if $\alpha = 0.10$, then in order to have reasonable computer run times, the approximate restrictions on the total sample size $k_1 + k_2$ are: for 10 distinct values, about 50; for 8 distinct values, about 80; for 6 distinct values, about 150; and for 5 distinct values, about 250 [1296, p. 664].[50]

4.18 Baker–Hubert and Ordering Theory

In 1977 Baker and Hubert published an article on inference procedures for ordering theory [53]. Given N observations on a set of n dichotomously scored test items representing certain skills or tasks, Baker and Hubert utilized ordering theory to identify a hierarchical organization among the n items. A directed graph representation of the n items motivated the ordering theory. Noting that it is relatively easy to extract a directed graph representation for a set of items, they emphasized that the problem remained as to how to assess whether this representation corresponds to the researcher's a priori notions regarding the hierarchy.

Observing that $n!$ possible enumerations of the n items is often too large for a practical permutation test, Baker and Hubert considered two alternatives. First, using formulae first given by Mantel [881] and Mantel and Valand [889], they obtained the exact mean and variance of the $n!$ possible enumerations of the n items

[50] In 1977 the CDC 6400 and the IBM System/370 were the dominant mainframe computers and the IBM 5100 was introduced in 1975 as the first portable computer, although it weighed nearly 50 pounds. By today's standards, these computers lacked both memory and speed, thereby severely limiting the calculation of exact tests. It was an IBM 370, Model 168, at the University of Illinois that Kenneth Appel and Wolfgang Haken used to solve the four-color map problem, which was published in 1977 in *Illinois Journal of Mathematics*. The "Four-color Conjecture" had stood unsolved for over 100 years and the proof that a flat map could be colored with just four colors so that contiguous countries would have different colors took 1,200 h of dedicated computer time on the University of Illinois IBM 370 mainframe.

and calculated a probability value based on a normal approximation. For the second alternative, they considered 1,000 random enumerations and used these to estimate the exact mean and variance. It should be noted that Baker and Hubert chose not to determine a resampling-approximation probability value directly from the 1,000 random enumerations, but used them to estimate a distribution. Unfortunately, a year later in 1978, the approach based on the mean and variance by Mantel and Valand was largely discredited [935]; see for example, two papers by Mielke in 1978 and 1979 [935, 936].

4.19 Green and Two Permutation Tests for Location

Also in 1977, Bert Green published an interactive FORTRAN computer program for one- and two-sample permutation tests of location [548]. Noting that Fisher's permutation tests of location had been described by Bradley in 1968 as "stunningly efficient" but "dismally impractical" [201], Green proposed a practical permutation program that contained two heuristics that permitted most of the permutations to be counted implicitly rather than explicitly. Both exact and resampling-approximation procedures were provided in the program.

Consider the two-sample case where n_1 and n_2 denote the numbers in the two samples with $n_1 \leq n_2$ and $n_1 + n_2 = n$, and let s_1 and s_2 denote the sums of the values in the two samples, respectively. Green's program used s_1 as its test statistic as s_1 is monotonic with the mean difference [548, p. 38]. Following Green, first the n values are ordered from least to greatest and the combinations of values are examined in lexicographic order. For example, let $n_1 = 3, n_2 = 5, n = 8$, and let $\{ijk\}$ signify $x_i + x_j + x_k$. Then the order is $\{1\,2\,3\}$, $\{1\,2\,4\}, \ldots, \{1\,2\,8\}$, $\{1\,3\,4\}, \ldots, \{1\,7\,8\}$, $\{2\,3\,4\}, \ldots, \{6\,7\,8\}$. The program tracks partial sums, where each sum is obtained by adding only one value to a partial sum. For example, in examining $\{1\,4\,5\}$, $\{1\,4\,6\}$, $\{1\,4\,7\}$, and $\{1\,4\,8\}$, $x_1 + x_4$ is computed only once, permitting the sequence of sums to be obtained with little effort.

Second, a simple heuristic permits the counting algorithm to skip many of the larger sums. Consider again $n_1 = 3$, $n_2 = 5$, and $n = 8$. Suppose $\{1\ 5\ 7\} > s_1$, then since $x_8 > x_7$, $\{1\ 5\ 8\} > \{1\ 5\ 7\}$, so $\{1\ 5\ 8\} > s_1$ and $\{1\ 5\ 8\}$ need not be computed. For another example, note that if $\{2\ 4\ 5\}$ exceeds s_1, then so will all further triples beginning with 2, i.e., $\{2\,4\,6\}$, $\{2\,4\,7\}, \ldots, \{2\,7\,8\}$. While this heuristic eliminates combinations of large numbers, Green introduced a second heuristic to eliminate combinations of small numbers. Suppose $n_1 = 5$, $n_2 = 6$, and $n = 11$; the program starts its enumeration with $\{1\ 2\ 3\ 4\ 5\}$. Now, if $\{1\ 8\ 9\ 10\ 11\} < s_1$, then all $\binom{10}{4} = 210$ combinations of $\{1\ * * * *\}$ will be less than s_1 and need not be examined further. On the other hand, supposed that $\{2\ 8\ 9\ 10\ 11\} \geq s_1$, then all combinations beginning with 2 must be examined. But if $\{2\ 3\ 9\ 10\ 11\} < s_1$, all of the $\binom{8}{3} = 56$ combinations of $\{2\ 3\ * * *\}$ must be less than s_1, so the program jumps to $\{2\ 4\ * * *\}$. As Green concluded, putting the two heuristics together permitted a very fast program and with $n_1 = n_2 \geq 10$, the amount of savings was over 90 % of the counts.

4.20 Agresti–Wackerly–Boyett and Approximate Tests

In 1979 Agresti, Wackerly, and Boyett suggested a new permutation procedure for $r \times c$ contingency tables in which only a random sample of all possible cell frequency configurations was analyzed [8]. The procedure was based on an innovative resampling algorithm by Boyett [199] (q.v. page 271).

A. Agresti

Alan Agresti earned his B.A. degree in mathematics from the University of Rochester in 1968 and his Ph.D. in statistics from the University of Wisconsin in 1972. His first position was in the Department of Statistics at the University of Florida in 1972 where he remained until his retirement in 2010. At the time of this writing, Agresti is Distinguished Professor Emeritus at the University of Florida. He has enjoyed visiting professor positions at Imperial College, London, Harvard University, the London School of Economics, and shorter visiting positions at the University of Florence and the University of Padova in Italy, Hasselt University in Belgium, Université Paris Diderot (Paris VII), Boston University, and Oregon State University.

D. Wackerly

Dennis Wackerly earned his B.S. degree from the University of Dayton in mathematics and computer science in 1967 and his M.S. and Ph.D. degrees in statistics from Florida State University in 1969 and 1973, respectively. His first position was in the Department of Statistics at the University of Florida at Gainesville where he remained for his entire career, retiring in 2007.

J.M. Boyett

James M. Boyett earned his B.S. degree in electrical engineering from the Georgia Institute of Technology in Atlanta in 1966, his M.A. degree from the University of Alabama at Huntsville in 1970, his M.S. degree from Michigan State University in 1971, and his Ph.D. in statistics and probability from Michigan State University in Lansing in 1974. Boyett served in many positions during his career. He was appointed Assistant Professor in the Department of Statistics at the University of Florida at Gainesville in 1974. In 1980 he was an Associate Research Scientist at the University of Florida; in 1986 he became Head of the Cancer Section, Department of Biostatistics and Epidemiology, The Cleveland Clinic Foundation; and in 1992 Professor,

(continued)

Department of Preventive Medicine, Division of Biostatistics and Epidemiology, University of Tennessee at Memphis, and Director, Biostatistics Shared Resource for the St. Jude Cancer Center.

In 1979 Agresti, Wackerly, and Boyett proposed a permutation procedure for approximating attained significance levels of exact conditional tests for $r \times c$ contingency tables [8]. They noted that in practice, contingency tables in which the expected cell frequencies are too small to employ asymptotic sampling distributions often occur. In these cases, they suggested an exact test of independence, such as Fisher's exact probability test, conditional on the observed marginal frequency totals. However, they observed that the number of tables, given the fixed marginal frequency totals, was often very large, making exact tests "infeasible." They then proposed to randomly generate a sufficient number of tables so that the attained significance level of the test could be estimated as accurately as was practically necessary.

Specifically, they suggested that a random sample of M distinct tables from all possible tables, S, could be achieved by repeating a procedure to generate random contingency tables M times. Then, after each table was generated, the desired test statistic would be calculated and its value compared to the value of the statistic for the observed table. As they explained, if the values of the statistics for X of the sampled tables provided at least as much evidence in favor of the alternative hypothesis as the value of the statistic for the observed table, then the estimated exact conditional level, $\hat{\alpha}$, was simply X/M. They specified a procedure to estimate α to within 0.01 with 99 % confidence as follows.

Since the number of test statistic values, X, in the tail of the distribution is a binomially-distributed variable with M trials and success probability α, then for large M, $\hat{\alpha} = X/M$ is approximately normally-distributed with mean α and variance $(1-\alpha)\alpha/M$. Thus, to estimate α within B units with $(1-\delta)100\,\%$ confidence requires

$$M \doteq \frac{(Z_{\delta/2})^2}{B^2}(\alpha)(1-\alpha) \ ,$$

where Z_δ denotes the $(1-\delta)$th quantile of the standard normal distribution. Since $\alpha(1-\alpha) \leq \frac{1}{4}$ for all α,

$$M \geq \frac{1}{4} \times \frac{\left(Z_{\delta/2}\right)^2}{B^2}$$

is sufficient for any α.

For example, to estimate α to within 0.01 with 99 % confidence,

$$M \geq \frac{1}{4} \times \frac{(2.576)^2}{(0.01)^2} = \left[\frac{2.576}{(2)(0.01)} \right]^2 = 16{,}589.44 \ .$$

Thus, Agresti, Wackerly, and Boyett showed that $M = 17{,}000$ was sufficient to estimate α to within 0.01 with 99 % confidence.[51]

4.21 Boyett and Random *R* by *C* Tables

In 1979 James M. Boyett (q.v. page 269) published an algorithm and associated FORTRAN subroutine RCONT to generate random $r \times c$ contingency tables with given fixed row and column marginal frequency totals [199].[52] Consider an $r \times c$ contingency table with fixed marginal frequency totals given by $a_{i.}$, $1 \leq i \leq r$ and $a_{.j}$, $1 \leq j \leq c$, and let

$$N = \sum_{i=1}^{r} a_{i.} = \sum_{j=1}^{c} a_{.j} \ .$$

First, employing a uniform pseudorandom number generator and a shuffling routine, Boyett generated a random permutation of the first N integers, x_1, $x_2, \ldots, x_N$, then partitioned the permuted integers into r groups of the row variable with each group S_i containing $a_{i.}$ values, $i = 1, \ldots, r$. For the column variable, the first N integers (not permuted) were partitioned into c groups with each group T_j containing $a_{.j}$ values, $j = 1, \ldots, c$. Thus, $S_1 = \{x_1, \ldots, x_{a_{1.}}\}$, $S_2 = \{x_{a_{1.}+1}, \ldots, x_{a_{1.}+a_{2.}}\}, \ldots,$ $S_r = \{x_{N-a_{r.}+1}, \ldots, x_N\}$, and $T_1 = \{1, \ldots, a_{.1}\}$, $T_2 = a_{.1} + 1, \ldots, a_{.1} + a_{.2}\}, \ldots,$ $T_c = \{N - a_{.c} + 1, \ldots, N\}$.

Then, the number of S_i values matching values in T_j yielded a_{ij} of one random $r \times c$ contingency table for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. In this manner, a random $r \times c$ contingency table was generated for each call of subroutine RCONT. However, subroutine RCONT lacked efficiency as it required N^2 attempted matches to generate the N cell frequencies for each random $r \times c$ contingency table.

For an example, consider a two-way contingency table in which both the row and column variables have three levels, $a_{1.}$, $a_{2.}$, and $a_{3.}$, and $a_{.1}$, $a_{.2}$, and $a_{.3}$, respectively. Finally, let the row marginal totals be $a_{1.} = 3$, $a_{2.} = 4$, $a_{3.} = 5$, and let the column marginal totals be $a_{.1} = 2$, $a_{.2} = 4$, $a_{.3} = 6$, for a total of $N = 12$ observations. On the first call to subroutine RCONT, a random permutation of the first $N = 12$ integers for S_i and the non-permuted first N integers for T_j might be:

[51] There is a mistake in the formula for M in Agresti, Wackerly, and Boyett [8, p. 78]. It has been corrected here.

[52] This was the algorithm that was employed by Agresti, Wackerly, and Boyett in their 1979 article on approximations of attained significance levels for $r \times c$ contingency tables [8].

Fig. 4.9 Random 3×3 contingency table from the data in S_i and T_j for $i, j = 1, 2, 3$

0	2	1	3
1	1	2	4
1	1	3	5
2	4	6	12

$$S_1 = \{3,\ 6,\ 9\}, \quad S_2 = \{12,\ 2,\ 5,\ 8\}, \quad S_3 = \{11,\ 1,\ 4,\ 7,\ 10\}\,,$$

$$T_1 = \{1,\ 2\}, \quad T_2 = \{3,\ 4,\ 5,\ 6\}, \quad T_3 = \{7,\ 8,\ 9,\ 10,\ 11,\ 12\}\,.$$

Then $a_{11} = 0$, as no values in S_1 match values in T_1; $a_{12} = 2$, as two values (3 and 6) in S_1 match values in T_2; $a_{13} = 1$, as only one value (9) in S_1 matches a value in T_3; $a_{21} = 1$, as only one value (2) in S_2 matches a value in T_1; $a_{22} = 1$, as only one value (5) in S_2 matches a value in T_2; $a_{23} = 2$, as two values (8 and 12) in S_2 match values in T_3; $a_{31} = 1$, as only one value (1) in S_3 matches a value in T_1; $a_{32} = 1$, as only one value (4) in S_3 matches a value in T_2; and $a_{33} = 3$, as three values (7, 10, and 11) in S_3 match values in T_3. The resulting random 3×3 contingency table would therefore be as shown in Fig. 4.9. The procedure would be repeated as many times as necessary with a random permutation of the first N integers partitioned into S_i, $i = 1, \ldots, r$, generated for each call to subroutine RCONT.

4.22 Looking Ahead

Early in the period from 1960 to 1979 non-computer methods were developed to generate random permutation sequences using tables of random integers. As the development of computers progressed, non-computer methods were superseded by computer-intensive methods for the generation of permutation sequences, resulting in vastly improved enumeration of permutation sequences. Random permutation sequences based on Monte Carlo procedures for both univariate and multivariate data structures soon followed. Thus, later in the period attention was largely focused on designing permutation versions of existing statistics. However, in 1976 Mielke, Berry, and Johnson introduced multi-response permutation procedures, which were designed specifically for data-dependent methods per se, in contrast to permutation alternatives to existing tests [971]. All this was made possible by the development and widespread availability of mainframe computers and user-friendly programming languages such as BASIC and FORTRAN. The mass marketing of personal computers was only just beginning in this period.

Permutation methods may be said to have "arrived" in the period from 1980 to 2000. While much of the permutation literature between 1960 and 1979 appeared in computer science magazines and journals, there was a dramatic shift away from computer science journals in the period from 1980 to 2000 and into statistical journals, as well as into medicine, psychology, public health, environmental science, biology, economics, ecology, and atmospheric science journals. A second shift was away from the use of moment-approximation probability values by calculating the exact moments of known distributions, such as the beta and Pearson type III distributions, to computing exact and resampling-approximation permutation probability values. This movement was facilitated by greatly improved computer speeds and the ready availability of desktop computers.

Of all the fields to embrace permutation methods in the 1960s and 1970s, psychology perhaps stands out, due to the large number of articles published in psychology journals on permutation methods in this period. Many early advances in statistics were made by psychologists, and others writing for psychologists. As Stephan Stigler wrote in an opening paragraph to a chapter on "Statistical Concepts in Psychology" in *Statistics on the Table: The History of Statistical Concepts and Methods*:

> [s]tatistics and psychology have long enjoyed an unusually close relationship — indeed more than just close, for they are inextricably bound together. That tie is of an unusual nature, with historical roots in the nineteenth century, and an understanding of this peculiar historical relationship can lead to a deeper understanding of contemporary applications [1321, p. 189].[53]

The intellectual climate in psychology, especially in experimental psychology, was amenable to the development and implementation of permutation methods for three reasons. First, the contributions of Eugene S. Edgington at the University of Calgary to permutation statistical methods in the period between 1960 and 1979 did much to promote permutation statistical methods in psychology. Psychologists took to permutation methods in large part because of the publication of Edgington's book on *Randomization Tests* in 1980, followed by a second edition in 1987, a third edition in 1995, and a fourth edition, co-authored with Patrick Onghena, in 2007. The initial book and subsequent editions were written by psychologists, for psychologists, and contained explicit examples based on psychological research, accompanied by associated computer routines.

Second, unlike other social sciences, such as political science, history, and sociology, psychology had a long history of developing quantitative methods especially attuned to its particular concepts and constructs; witness such illustrious psychometricians as Bernard Babington Smith, Cyril Burt, Jacob Cohen, Clyde Coombs, Leon Festinger, J. Paul Guilford, William Hays, Clark Hull, Everett Lindquist, Quinn McNemar, Sidney Siegel, Charles Spearman, S. Smith Stevens, Edward Thorndike, Louis Thurstone, and John W. Whitfield. In addition, psychology provided numerous outlets for the publication of quantitative methods with journals such as *Psychometrika*, *Educational and Psychological Measurement*, *Applied Psychological Measurement*, *Psychological Methods*, *Psychological*

[53] The American Psychological Association has long had a division devoted to quantitative/statistical psychology (Division 5—Evaluation, Measurement and Statistics), which was one of the Charter Divisions of the APA, and in the spring of 2012 the American Statistical Association announced a new section on Statistics and Measurement in Psychology and Education [715, p. 9].

Table 4.6 Summary statistics Q_1, Q_2, and Q_3 based on N studies for four surveys conducted in 1955, 1977, 1995, and 2006

	Holmes		Marszalek et al.	
Statistic	1955	1977	1995	2006
N	448	507	527	690
Q_1	25.50	18.13	14.00	18.00
Q_2	59.95	48.40	32.00	40.00
Q_3	131.30	94.00	87.50	136.00

Reports, *British Journal of Mathematical and Statistical Psychology*, and *Behavior Research Methods*.[54,55]

Third, in general psychologists study small nonrandom samples, for which exact permutation methods are ideally suited. In 1965, in a survey of eleven major American psychology journals that published original research, Dukes reported that between 1939 and 1963 there were 246 published experiments with only a single subject [362],[56] and Cochrane and Duffy, in a 1974 examination of two British psychology journals, found that all the studies in those two journals used 25 or fewer subjects [262].

Cooper Holmes and his co-authors have repeatedly investigated sample sizes employed in psychological experiments [643–645, 896]. A summary of his findings and those of Marszalek, Barber, Kohlhart, and Holmes [896] are provided in Table 4.6, where the findings are categorized by the lower quartile (Q_1), the median (Q_2), and the upper quartile (Q_3), of the number of subjects studied in 1955, 1977, 1995, and 2006. The results in Table 4.6 indicate not only that psychologists often study small samples, but also that the use of small samples has changed very little over the span of the 52 years surveyed.

Finally, in a study that examined journals in four sub-areas of psychology (abnormal, developmental, applied, and experimental), Holmes found the experimental area used the fewest number of subjects [644]. In a review of $N = 161$ studies in experimental psychology journals in 1977, Holmes calculated Q_1 to be 7.51, Q_2 to be 12.20, and Q_3 to be only 31.68 [644].

[54] *Behavior Research Methods* was published as *Behavior Research Methods & Instrumentation* from 1969 to 1983 and as *Behavior Research Methods, Instruments, & Computers* from 1984 to 2004.

[55] A historical account of the development of statistics in psychology between 1925 and 1950, with an emphasis on the analysis of variance, is provided in a 1980 article by Rucci and Tweney [1205].

[56] For experimental studies based on only one subject, Edgington provides an interesting justification in an article on "Statistical inference from $N = 1$ experiments" published in 1967 in *Journal of Psychology* [389].

5 1980–2000

The 1960s and 1970s witnessed a profusion of published algorithms and programs designed to generate permutation sequences with speed and efficiency, beginning with a non-computer procedure by C.R. Rao in 1961 that used a table of uniform random numbers [1154], and the first explicit description of a computer algorithm by Tompkins in 1956 [1364]. In addition, Ord-Smith [1068, 1069], Rabinowitz and Berenson [1149], and Sedgewick [1242] provided extensive summaries of the literature on the generation of permutation sequences in this period. As late as 1989, Eric Noreen observed in reference to permutation methods:

> [t]he next few years are likely to be an exciting period for those involved in testing hypotheses. Recent dramatic decreases in the costs of computing now make revolutionary methods for testing hypotheses available to any one with access to a personal computer. These [permutation] methods are easy to understand, very general, and can often avoid troublesome assumptions that are required with conventional methods [1041, p. 1].

Nonetheless, Noreen noted that "[s]ince exact randomization tests are seldom feasible, this book will henceforth be concerned only with approximate randomization [resampling-approximation] tests" [1041, p. 15].

Progress on the development of permutation methods continued unabated during the 1980s and 1990s, paralleling advancements in high-speed computing and the subsequent wide-spread availability of both university mainframes and, later in the period, personal desktop computers. Also, a number of books were published in this period that introduced permutation methods to a wide variety of audiences, accompanied by a decided shift in the literature away from the computer science journals that had focused on issues of efficiently calculating permutation sequences and into discipline journals that were more focused on permutation statistical tests. These progressions were accompanied by an increasing emphasis on statistical applications of permutation methods, both exact and resampling, since efficient computer-based permutation-sequence generators were widely available.

A brief overview of the development of permutation statistical methods in the period from 1980 to 2000 introduces this chapter and is followed by an in-depth treatment of selected contributions, both statistical and computational. The chapter

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9_5,
© Springer International Publishing Switzerland 2014

concludes with a look ahead at the continuing development of permutation statistical methods beyond the year 2000.

5.1 Overview of This Chapter

Permutation statistical methods arrived at a new level of maturity between 1980 and 2000, primarily as a result of two factors: (1) greatly improved computer clock speeds and (2) widely-available desktop computers. Boardman, in a 1984 paper, for example, discussed the impact of smaller computers on statistical data analysis at that time [174]. While interest continued in the study of linear rank-order statistics [987], the same period witnessed a dramatic shift in sources of permutation publications. In the previous period, from 1960 to 1979, nearly all published papers on permutation methods appeared in computer journals, such as *Communications of the ACM*, *The Computer Bulletin*, *ACM Transactions on Mathematical Software*, and *The Computer Journal*. However, in the period 1980 to 2000 there was a shift away from computer journals and into statistical journals, such as *Biometrika*, *Journal of the American Statistical Association*, *The American Statistician*, *Communications in Statistics*, and *Applied Statistics*.[1] An even more dramatic change occurred in this period as an increasing number of published papers on permutation statistical methods began appearing in discipline journals, such as *American Journal of Public Health*, *Educational and Psychological Measurement*, *Psychometrika*, *Econometrica*, *Ecology*, *Behavior Research Methods, Instruments, & Computers*, *Journal of Applied Meteorology*, and *Vegetatio*.[2]

In addition, a number of books on permutation methods appeared in this period, beginning with the first edition of Edgington's *Randomization Tests* in 1980 [392], a second edition seven years later in 1987 [393], and a third edition in 1995 [394]. Edgington's book was quickly followed by Hubert's *Assignment Methods in Combinatorial Data Analysis* in 1987 [666]; Noreen's *Computer Intensive Methods for Testing Hypotheses* in 1989 [1041]; Westfall and Young's *Resampling-based Multiple Testing* in 1993 [1437]; Good's *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* [523] and *Permutation, Parametric and Bootstrap Tests of Hypotheses* in 1994 [522]; Manly's first edition of *Randomization and Monte Carlo Methods in Biology* in 1991 [875], followed by a second edition in 1997 [876]; Weerahandi's *Exact Statistical Methods for Data Analysis* in 1995 [1421]; Simon's *Resampling: The New Statistics* in 1997 [1277]; Good's *Resampling Methods: A Practical Guide to Data Analysis* in 1999 [524]; and Lunneborg's *Data Analysis by Resampling: Concepts and Applications* in 2000 [858].

[1] Continued by *Journal of the Royal Statistical Society, Series C*.

[2] Continued by *Plant Ecology*.

5.2 Development of Computing

The Apple I personal computer (PC) was introduced in 1976 and consisted of a limited production circuit board for electronic hobbyists; consequently, only 220 Apple I personal computers were sold.[3] The Apple II personal computer was introduced in 1977 and various models ending with the Apple IIc Plus followed through 1993. In contrast to the Apple I, over six million Apple II computers were sold. The Macintosh personal computer was introduced in 1984 and consisted of a small screen, a keyboard, a mouse with one button, and a graphical user-interface (GUI). It is estimated that through 2011 some 20 million Macintosh personal computers were sold.

On 12 August 1981 the first IBM PC was introduced as model 5150. It ran on a 4.77 MHz Intel 8088 microprocessor and came with 16 kilobytes of memory and an optional color monitor. In April of 1982 the GRiD Compass 1101 was introduced as the first laptop computer; price: $8,150. It ran an Intel 8086 processor at 8 MHz, and had 340 kilobytes of magnetic bubble memory. The GRiD Compass, designed by William Grant Moggridge, had a clamshell case, roughly 15 × 22 in., which opened to reveal a luminous screen on top that folded over the keyboard on the bottom. It was a GRiD Compass that was used by astronaut John Creighton on the Space Shuttle Discovery in 1985. In 1983 Compaq Computer Corporation marketed the first PC clone that was 100 % compatible with IBM's PC; first year sales: $111 million. Also in 1983, the IBM 5160 (or simply, the IBM XT) personal computer was released running an Intel 8088 processor at 4.77 MHz. It was quickly followed by the IBM 286 XT in 1986 running an Intel 80286 processor at 6 MHz; the IBM 386 SLC in 1991 with available Intel 80386 processors running at 16, 20, and 25 MHz[4]; and the IBM 486 SLC in 1992 with available Intel 80486 processors running at either 50 or 66 MHz.

In 1984 John Ashworth Nelder published a paper on the present position and potential developments of statistical computing [1027]. As such, the paper provides a window into statistical computing in the early 1980s. Nelder provided a brief history of computing in statistics, from the Analytical Engine of Charles Babbage to the electronic computers of the time. He then discussed the present state of statistical computing with particular emphasis on computing algorithms, hardware, computing languages, control languages, and operating systems. Nelder concluded the paper by predicting that future developments in statistics over the next 150 years must involve the computer [1027]. As Brian Edward Cooper noted, what Nelder neglected

[3] On 8 August 2013 a retired school psychologist from Sacramento, California, sold one of the few remaining original Apple I computers (Serial Number 01-0025) that had been assembled by hand by Steve Wozniak in Steve Jobs parents' garage. It fetched $387,750 at Christie's Auction House; original price in 1976: $666.66. When the Apple II personal computer was introduced in 1977, customers were allowed to trade in their Apple I computers, making surviving Apple I computers very rare.

[4] The 386 SLC was known inside IBM at the Super Little Chip for its initials.

to forecast was (1) the future importance of micro-computers, both as stand-alone machines and as intelligent terminals within a network, and (2) the designing of operating systems that catered to the naïve rather than the expert user [276].

Also in 1984, two economists living in Santa Monica, California, William W. Gould and William H. Rogers initiated the development of Stata, a comprehensive (eventually) statistical analysis package. Stata[5] began as a small executable module written in C, supplemented by a number of higher-level language programs written in Stata's proprietary language [618]. S is a statistical computing language developed by John Chambers and Trevor Hastie, along with Richard A. Becker, Alan Wilks, and William S. Cleveland, at Bell Laboratories around 1975–1976. In 1990 an S clone called R was developed by Robert Gentleman and Ross Ihaka at the University of Auckland, New Zealand, who were looking for a statistical environment to use in their teaching laboratories. At the time, the laboratories were populated with Macintosh computers and they knew of no suitable statistical software available for the Macintosh environment [1422]. Initial versions of R were provided to Carnegie Mellon University and the user feedback indicated a positive reception for the new language. In June of 1995 R was released as freeware, about the same time that Martin Mächler joined the development team, and in 2000 John Chambers, one of the original developers of S, joined the core team of R. At present, R is one of the most popular programming environments in use [618].

In 1986 Eddy, Huber, McClure, Moore, Stuetzle, and Thisted provided a snapshot of the present and future needs of computer equipment and operating expenses for computing facilities to support statistical research [373]. Their published article was actually a report on a workshop on the "Use of Computers in Statistical Research" held at Carnegie Mellon University in Pittsburgh, Pennsylvania, and sponsored by a grant from the Mathematical Sciences Division of the Office of Naval Research (ONR) to the Institute of Mathematical Statistics (IMS). The comprehensive report surveyed 30 universities and painted a picture of computing needs at major universities in the middle 1980s, ranging from hardware and software to physical plants and support staff.

On 12 June 2005, a 50-year-old Steven Paul (Steve) Jobs spoke to a group of students at Stanford University, recalling his campus days at a "lesser institution"—Reed College in Portland, Oregon. Throughout the Reed College campus, he remembered, every poster, every label on every drawer, was beautifully done in hand calligraphy. Because Jobs had dropped out of Reed College and therefore had no required classes, he elected to enroll in a calligraphy class. There he learned about fonts, typefaces, kerning, tracking, leading, serifs, ligatures, and all that makes typography interesting.[6] Ten years later, Jobs designed the Macintosh personal computer and introduced something unprecedented at the time—a wide variety of computer fonts. These fonts included Times New Roman and Helvetica, but also

[5] Stata is a portmanteau of the words "statistics" and "data."

[6] On this topic, see a wonderful little book published in 2011 by Simon Garfield titled *Just My Type: A Book About Fonts* [495].

several designed by Jobs, which he named after cities he loved, such as Chicago, Toronto, Venice, and Los Angeles [495, pp. 1–2].

During this period, work also continued on improving the computational efficiency of permutation tests, inspired by the ease of calculations due to dramatic increases in computer speed and storage. Between 1980 and 1999 a number of "algorithmic tricks" were developed that substantially reduced computing time for many permutation statistical methods [489, 1397]. Early in the period, Berry published a highly efficient algorithm to generate permutations of multi-sets in Gray-code order [108]. Later in the period, Balmer [56] and Dallal [311] utilized recursive routines to efficiently generate both statistics and probability values, and Thakur, Berry, and Mielke [1349], Berry and Mielke [126, 131, 141], and Berry, Mielke, and Helmericks [158, 159] enhanced the recursion procedure by coupling recursive routines with the use of an arbitrary initial value to initiate the recursion. A second algorithmic innovation was to recognize that only the variable portion of a statistical formula needed to be computed for each permutation, thereby increasing the efficiency of exact permutation tests. But, by far, the most important innovation was the introduction of a highly efficient network algorithm by Mehta and Patel in 1980 and 1983 [919, 920].

At the beginning of this period in 1980, Cyrus Mehta and Nitin Patel [919] introduced a network algorithm that proved to be a highly efficient method for calculating exact permutation tests. Originally designed for computing exact tests for $2 \times c$ contingency tables, the algorithm was quickly extended to the more general problem of $r \times c$ contingency tables by Pagano and Taylor Halvorsen in 1981 [1081] and by Mehta and Patel in 1983 [920]. Interest continued in this period on computational methods for both exact and resampling analyses of $r \times c$ contingency tables with articles by Balmer [56]; Romesburg, Marshall, and Mauk [1191]; Phillips [1125]; Berry and Mielke [126, 129–131, 134]; Kannemann [709, 710]; Zar [1486]; Pagano and Taylor Halvorsen [1081]; Patefield [1089]; Saunders [1223]; Mielke and Berry [947, 949]; and Baglivo, Olivier, and Pagano [45]. Extensions to multidimensional (r-way) contingency tables were provided by Kreiner [771]; Mielke and Berry [948, 953, 955]; Berry and Mielke [136]; Mielke, Berry, and Zelterman [983]; and Zelterman, Chan, and Mielke [1489].

In the period between 1980 and 2000, permutation tests branched out from their home in statistics to include a variety of other disciplines, most notably in psychology with articles by Berry and Mielke [121, 122, 127, 134] and Mielke and Berry [946, 954, 955]; pharmacology and physiology with an important article by Ludbrook [849]; biomedical sciences with articles by Ludbrook and Dudley [856], Dallal [311], and Zimmerman [1495, 1496]; anthropology with articles by Mielke, Berry, and Eighmy [970] and Berry, Mielke, and Kvamme [161]; ecology with articles by Zimmerman, Goetz, and Mielke [1494] and Biondini, Mielke and Berry [166]; wood science with an article by Pellicane, Potter, and Mielke [1116]; geoscience with an article by Romesburg [1190]; and atmospheric science with articles by Mielke, Berry, and Brier [968], Gray, Landsea, Mielke, and Berry [547], Mielke [938, 939], Mielke, Berry, and Medina [981], Wong, Chidambaram, and Mielke [1468], Tucker, Mielke, and Reiter [1374], Lee, Pielke, and Mielke [809], Kelly, Vonder Haar, and Mielke [716], Cotton, Thompson, and Mielke [288], and Mielke, Berry, Landsea, and Gray [979, 980].

While many of the contributions to the permutation literature during this period concentrated on efficient means for calculating permutation versions of existing statistics, advancements in computational efficiency allowed for the development of a wider variety of statistical tests, tailored to the specific requirements of whatever problem was under consideration at the time. Consequently, a few researchers utilized permutation structures to develop new statistical measures and tests.

Permutation versions of existing statistics included Fisher's exact probability test by Verbeek and Kroonenberg [1397], Berry and Mielke [121, 127, 131], Mehta and Patel [920–922], Mielke and Berry [949], Baglivo, Olivier, and Pagano [45], Joe [688], and Zar [1486]; analysis of variance in its various forms by Manly and Francis [878] and Routledge [1198]; the chi-squared test of independence by Mielke and Berry [947], Baglivo, Olivier, and Pagano [45], and Romesburg, Marshall, and Mauk [1191]; various goodness-of-fit tests by Baglivo, Olivier, and Pagano [45], Mielke and Berry [950], and Tritchler [1370]; the Kolmogorov–Smirnov test by Romesburg, Marshall, and Mauk [1191]; the Terpstra–Jonckheere test for ordered alternatives by Berry and Mielke [145] and Mielke and Berry [960]; Hotelling's generalized T^2 statistic by Blair, Higgins, Karniski, and Kromrey [169] and Mielke, Berry, and Neidt [982]; and the Wilcoxon signed-ranks test by Dallal [311] and Zimmerman [1495].

Also, the Wilcoxon–Mann–Whitney two-sample rank-sum test by Dallal [311], Zimmerman [1496], and Berry and Mielke [155]; the likelihood-ratio test by Baglivo, Olivier, and Pagano [45]; one-way analysis of variance by Berry and Mielke [121]; the odds-ratio by Vollset and Hirji [1399] and Vollset, Hirji, and Elashoff [1400]; the Goodman–Kruskal τ_b measure of nominal contingency by Berry and Mielke [126, 135]; Cohen's kappa measure of agreement by Berry and Mielke [133]; Cochran's Q test by Mielke and Berry [952, 954] and Berry and Mielke [143]; logistic regression by Hirji, Mehta, and Patel [631] and Tritchler [1370]; partial regression coefficients by Anderson and Legendre [20]; various two-sample tests by Zimmerman [1495, 1496], Baker and Tilbury [54], Chen and Dunlap [250], and Edgington and Khuller [395]; the McNemar test by Baker and Tilbury [54]; survival analysis by Sun and Sherman [1335]; g-sample empirical coverage tests by Mielke and Yao [989, 990]; and the Cochran–Armitage test for trend by Mehta, Patel, and Senchaudhuri [924].

At the same time, Mielke and his collaborators focused their work on designing permutation tests, such as MRPP (q.v. page 254), that were not simply permutation versions of existing statistics. Conventional statistical tests and measures, both parametric and non-parametric, are based on squared Euclidean distances among data points. Examples include two-sample t tests, various F tests, ordinary least-squares (OLS) regression, and non-parametric tests such as the Wilcoxon–Mann–Whitney two-sample rank-sum test, the Kruskal–Wallis analysis of variance rank test, the Terpstra–Jonckheere test for ordered alternatives, and the Friedman two-way analysis of variance for ranks.

A Euclidean-distance function based on absolute distances among data points was incorporated into new permutation tests for matched-pairs designs by Mielke and Berry [946], Berry and Mielke [125, 142], Brockwell and Mielke [219], Mielke and Berry [945], and Mielke, Berry, and Neidt [982]; completely randomized designs by Mielke, Berry, and Brier [968], Berry, Kvamme, and Mielke [118, 119], Berry and Mielke [120, 123, 154], O'Reilly and Mielke [1070], Brockwell, Mielke, and Robinson [220], Mielke [941, 943], Mielke, Berry, Brockwell, and Williams [969], and Mielke and Berry [958]; and randomized block designs by Mielke [938], Tucker, Mielke, and Reiter [1374], Brockwell and Mielke [219], Mielke and Berry [945], Berry and Mielke [150], and Mielke and Iyer [984].

Also, contingency table analyses by Berry and Mielke [126, 130, 135, 136, 138, 139], Mielke [937], Mielke and Berry [948], and Zelterman, Chan, and Mielke [1489]; goodness-of-fit tests by Mielke and Berry [950] and Berry and Mielke [140]; spatial analysis by Reich, Mielke, and Hawksworth [1159]; multiple regression by Mielke and Berry [956, 957] and Berry and Mielke [149–151, 153, 154]; and measures of agreement and consensus by Berry and Mielke [133,137–139,144,146]. In addition to emphasizing the congruence between a data space and an ordinary Euclidean analysis space in three papers by Mielke [938, 939, 941], a number of detailed examples in the latter two papers [939, 941] suggested a major improvement in robustness for analyses based on ordinary Euclidean rather than squared Euclidean distances (q.v. page 404).

5.3 Permutation Methods and Contingency Tables

The period between 1980 and 2000 witnessed a continuation of the work done on contingency table analyses between 1960 and 1979.[7] In 1981 Patefield published an efficient method of generating random $r \times c$ contingency tables with fixed row and column marginal frequency totals [1089]. The Patefield FORTRAN subroutine, RCONT2, was designed to be an improvement over the previously published algorithm of Boyett, RCONT (q.v. page 271).

As Patefield explained, under the null hypothesis of no association between row and column categories, the joint probability distribution of a random table is given by a_{ij}, $i = 1, \ldots, r$ and $j = 1, \ldots, c$, conditional on the row and column totals, $a_{i.}$, $1 \le i \le r$ and $a_{.j}$, $1 \le j \le c$. Patefield considered the conditional distribution of a table entry a_{lm} given the table entries in previous rows, i.e., a_{ij}, $i = 1, \ldots, l-1$ and $j = 1, \ldots, c$, and the previous table entries in row l, i.e., a_{lj}, $j = 1, \ldots, m-1$.

Assuming valid conditioning table entries, the range of the conditional distribution is from a minimum of

$$\max\left\{0, a_{l.} - \sum_{j=1}^{m-1}\left[a_{lj} - \sum_{j=m+1}^{c}\left(a_{.j} - \sum_{i=1}^{l-1} a_{ij}\right)\right]\right\}$$

[7] For an excellent bibliography on contingency table analysis from 1900 to 1974, see a 1976 article by Killion and Zahn in *International Statistical Review* [754].

to a maximum of

$$\min\left[\left(a_{.m} - \sum_{i=1}^{l-1} a_{im}\right), \left(a_{l.} - \sum_{j=1}^{m-1} a_{lj}\right)\right].$$

The table entries a_{rm}, $m = 1, \ldots, c$, in the last row and a_{lc}, $l = 1, \ldots, r$, in the last column of the table were obtained by Patefield from the previous $(r - 1) \times (c - 1)$ table entries and the fixed row and column marginal frequency totals $a_{i.}$, $i = 1, \ldots, r$, and $a_{.j}$, $j = 1, \ldots, c$.

Patefield compared subroutine RCONT2 with Boyett's subroutine RCONT using contingency tables of sizes 2×7, 3×4, 4×4, 5×5, and 6×6 with sample sizes of $n =$ 10, 20, 30, 50, 100, 200, 500, and 1,000. The timings were based on 1,000 calls to the subroutines and for all tables the row marginal frequency totals were approximately equal to n/r and the column marginal frequency totals were approximately equal to n/c. Patefield concluded that whereas the time required to generate random tables using Boyett's RCONT algorithm was approximately proportional to sample size, subroutine RCONT2 was more dependent on the dimensions of the table [1089, p. 94].

Previously, in 1954, Goodman and Kruskal had introduced statistic gamma (γ) for ordered $r \times c$ contingency tables [534]. In 1981 Gans and Robertson considered the γ statistic for 2 × 2 contingency tables with small and moderate sample sizes [494].[8] For comparison, they also looked at Pearson's product-moment correlation coefficient, ρ; Spearman's rank-order correlation coefficient, ρ_s; and Kendall's measure of rank-order correlation, τ. Based on exact analyses of nine sets of cell probabilities, calculated for sample sizes of $n = 10, 30$, and 50, Gans and Robertson concluded that convergence to normality was much slower for γ than for ρ and, further, that the distribution of γ was much more irregular than for ρ.

In 1981 Romesburg, Marshall, and Mauk introduced FORTRAN program FITEST for computing an exact chi-squared goodness-of-fit test between an observed and a theoretical distribution [1191]. Echoing Radlow and Alf (q.v. page 249), they argued that chi-squared goodness-of-fit tests should be based on the following definition: a given table is as deviant or more deviant than the observed table if the calculated value of its chi-squared statistic is equal to or larger than the calculated value of the chi-squared statistic for the observed table [1191, p. 48].

Program FITEST was based on a more efficient enumeration algorithm than that presented by Radlow and Alf in 1975 [1150]. In addition, for larger problems where complete enumeration was not possible, program FITEST optionally provided for an approximate probability value using Monte Carlo resampling methods. Unusual for the time, Romesburg, Marshall, and Mauk also provided a complete listing of program FITEST, written in FORTRAN IV [1191, pp. 53–58].

[8] When Goodman and Kruskal's γ statistic is restricted to 2 × 2 contingency tables, it reduces to Yule's Q statistic, introduced in 1912 [1480, 1481].

In 1981 Pagano and Taylor Halvorsen presented an efficient algorithm for calculating the exact permutation significance value for $r \times c$ contingency tables [1081]. The primary feature of this new algorithm was a method to obtain exact probability values for $r \times c$ contingency tables without complete enumeration. In brief, they used a recursion routine beginning with the tail probability values of the tail in which the observed table fell, but terminated when the probability value of the observed table was reached. Then the other tail was examined, summing the tail probability values until a table was found with a larger probability value than the probability value of the observed table. They noted that further savings could be effected if the researcher was interested only in whether the observed table was or was not significant at some predetermined level, and not interested in the actual probability value [1081, p. 933].

In 1982 Phillips presented a simplified, but accurate, algorithm for the Fisher–Yates exact probability test for 2 × 2 contingency tables [1125]. Building on the recursion approach of Feldman and Klinger [424] (q.v. page 220), Phillips employed an arbitrary initial value for the recursion, summed all the resultant recursion values to obtain a total, then divided each recursion value by the total to obtain the probability values. The Phillips' algorithm was identical to the general recursion technique utilizing an arbitrary initial value described by Frank Yates in 1934 [1472, p. 219] (q.v. page 44).

In 1982 Kannemann published two articles in *Biometrical Journal* on the exact evaluation of $r \times c$ contingency tables [709, 710]. In these two articles he presented an algorithm that permitted the exact evaluation of sparse $r \times c$ contingency tables with fixed row and column marginal frequency totals, and further claimed that he had solved the long-standing associated enumeration problem of the number of possible tables containing integer arrays with fixed row and column marginal frequency totals [710]. Kannemann was quickly challenged by Kroonenberg and Verbeek as the claims of Kannemann lacked verisimilitude [774]. They pointed out that the algorithm proposed by Kannemann had previously been introduced by Hancock in 1975 [583], with suggested improvements by Howell and Gordon in 1976 [657] and by Cantor in 1979 [241]. Additionally, they noted, even more efficient algorithms by Boulton [185], Agresti and Wackerly [7], and Baker [55] had previously been published in 1974 and 1977.

Kroonenberg and Verbeek questioned the claim by Kannemann that he could analyze two-way contingency tables up to 50 × 50 and three-way contingency tables by complete enumeration, noting that the smallest family of 50 × 50 contingency tables with only $n = 50$ cases would require enumerating $50! > 10^{64}$ tables[9] and for three-dimensional contingency tables, the smallest families would be larger than the third power of families of two-dimensional contingency tables [774, pp. 719–720]. Finally, with respect to Kannemann's claim of closed formulae for the number of possible contingency tables given fixed row and column marginal frequency totals,

[9] Authors' note: the actual number is 30,414,093,201,713,378,043,612,608,166,064,768,844,377, 641,568,960,512,000,000,000,000 contingency tables.

Fig. 5.1 Notation for a standard 2 × 2 contingency table

x	$r-x$	r
$c-x$	$n-r-c+x$	$n-r$
c	$n-c$	n

Kroonenberg and Verbeek simply stated "they still do not exist," noting however that good approximation formulae had been developed by Boulton and Wallace [187], Good [518], and Gail and Mantel [490] (q.v. page 266).

In 1983 Berry and Mielke developed a rapid FORTRAN subroutine for the Fisher–Yates exact probability test [122]. Subroutine FEP (Fisher Exact Probability) utilized the hypergeometric distribution to calculate one- and two-sided exact probability values for the Fisher–Yates exact probability test. The approach of Berry and Mielke differed from the traditional approach exemplified by Robertson in 1960 [1174], Gregory in 1973 [553], Tritchler and Pedrini in 1975 [1371], and Bedeian and Armenakis in 1977 [91]. The conventional approach began with the premise that, since the Fisher–Yates exact probability test is comprised of nine factorial expressions for each table, it is time consuming and expensive to compute. Therefore, only those probability values less than or equal to the observed probability value should be calculated, as advocated by Howell and Gordon in 1976 [657], Fleishman in 1977 [466], and Romesburg, Marshall, and Mauk in 1981 [1191].

Subroutine FEP operated under a different premise. Consider a standard 2 × 2 contingency table, such as depicted in Fig. 5.1, consisting of n cases, with x denoting the observed frequency of any cell, and with r and c representing the row and column marginal frequency totals, respectively, of x. Then, the point-probability (P) value of any x is given by

$$P(x|n,r,c) = \frac{\dbinom{c}{x}\dbinom{n-c}{r-x}}{\dbinom{n}{r}}.$$

Solving the recursive relation $P(x+1|n,r,c) = P(x|n,r,c) \times f(x)$ for $f(x)$ yields

$$f(x) = \frac{(r-x)(c-x)}{(x+1)(n-r-c+x+1)},$$

which may be employed to enumerate the complete distribution of $P(x|n,r,c)$, $v \leq x \leq w$, where $v = \max(0, r+c-n)$ and $w = \min(r,c)$. The one-sided (P_1) and two-sided (P_2) probability values are then given by

$$P_1(x|n,r,c) = \sum_{k=v}^{w} I_k J_k P(k|n,r,c)$$

and

$$P_2(x|n,r,c) = \sum_{k=v}^{w} J_k P(k|n,r,c),$$

respectively, where

$$I_k = \begin{cases} 1 & \text{if } \operatorname{sgn}(kn - rc) = \operatorname{sgn}(xn - rc), \\ 0 & \text{otherwise,} \end{cases}$$

and

$$J_k = \begin{cases} 1 & \text{if } |kn - rc| \geq |xn - rc|, \\ 0 & \text{otherwise,} \end{cases}$$

for $k = v, \ldots, w$. Thus, the procedure enumerated all possible probability values using recursion for $v \leq x \leq w$, discarding those greater than the observed probability value, and summing the remaining probability values. Any recursion process requires an initial starting value. In this case Berry and Mielke used

$$P(v|n,r,c) = \exp\left[\ln\binom{c}{v} + \ln\binom{n-c}{r-v} - \ln\binom{n}{r}\right],$$

where the natural logarithms of the combinations were approximated by an expansion of Stirling's formula (q.v. page 227) to approximate factorial expressions given by

$$n! \doteq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n .$$

Berry and Mielke demonstrated that this recursion approach over the entire permutation distribution was much more efficient than the conventional approach.

5.4 Yates and 2 × 2 Contingency Tables

Frank Yates (q.v. page 37) presented a paper before the Royal Statistical Society on "Tests of significance for 2 × 2 contingency tables" in 1984, the golden anniversary of his classic 1934 article on "Contingency tables involving small numbers and the χ^2 test" published in *Journal of the Royal Statistical Society, Series A* in the same year [1476]. The continuity correction for chi-squared, first proposed by Yates in 1934, had received considerable attention over the intervening 50 years; see for example, articles by Plackett in 1964 [1135], Mantel and Greenhouse in 1968 [886], Conover in 1974 [271], Miettinen in 1974 [991], Haber in 1980 [564], Haviland in 1990 [599], Peritz in 1992 [1117], Martín Andrés, Herranz Tejedor, and Luna del

Castillo in 1992 [898], and Martín Andrés, Sánchez Quevedo, Tapia García, and Silva Mato in 2005 [903]. Although much of the 1984 paper, like the 1934 article, addressed the role of the continuity correction in analyzing 2 × 2 contingency tables and had little to do with permutation methods per se, Section 15 of the 1984 paper considered exact two-sided probability values for tests of 2 × 2 contingency tables.

The proper method for calculating two-sided probability values for the Fisher–Yates exact probability test has long been controversial, and is still being debated; see for example, articles by Cormack in 1986 [280] and Mantel in 1990 [885]. The debate began in 1941 when Edwin Wilson published 2 × 2 contingency table data on the potency of two viruses in the journal *Science* [1461] and R.A. Fisher responded with a letter to *Science* in which he provided a one-tailed test for Wilson's virus data [456].[10] David Finney (q.v. page 159), noticing that Wilson's original statement of the problem required a two-tailed test, wrote to Fisher and inquired as how to test the null hypothesis with a two-tailed test, adding that he would be "grateful for your [Fisher's] views." Fisher replied that he felt he could defend "the simple solution of doubling the total probability."[11] Yates took the position that "the rule for determining the two-sided probability, if this is required, should be *to double the observed one-tail probability*. This is invariant under transformation, whereas basing two-sided probabilities on equal but opposite deviations is not" [1476, p. 442].[12]

In this same article Yates also took Haber [564] to task for his investigation of corrected and uncorrected chi-squared statistics. In 1980, in a comprehensive analysis of 2 × 2 contingency tables, Haber had compared several corrected chi-squared tests, including that of Yates', with the uncorrected chi-squared test. Haber found that in two-sided tests with 2 × 2 contingency tables, Yates' corrected chi-squared test yielded probability values that were too high. Haber had defined exact two-sided probability values for chi-squared in 2×2 contingency tables as the sum of the probabilities of all values with deviations equal to or greater than that of the observed deviation, rather than simply doubling the exact one-tail probability value as Yates had recommended [564]. In his conclusion, Yates argued that the Fisher–Yates exact probability test for 2×2 contingency tables was the "only rational test, whether both, one, or neither of the margins are determined in advance" [1476, p. 446], making the Fisher–Yates exact probability test the gold standard for 2 × 2 contingency tables. Here, Yates was trying to determine what application of a chi-squared-type test could best approximate the results from an exact test.[13] Finally, Yates concluded

[10] In 1941 Edwin B. Wilson published a short note on "The controlled experiment and the four-fold table" in *Science* in which he explored the use of the chi-squared test statistic, chi-squared with Yates' correction, and Fisher's exact probability test. The analyses were based on data on the potency of two viruses injected into a small sample of six mice.

[11] Fisher, quoted in Yates [1476, p. 444].

[12] Emphasis in the original.

[13] This was in contrast to Starmer, Grizzle, and Sen who stated in 1974 "[t]here seems to be no good reason to use the exact test as the standard of comparison for competing tests" of association, instead suggesting as a gold standard a randomized version of the exact test [1315, p. 377].

that for the Fisher–Yates exact probability test, one-tail probability values should be used, but if a two-sided probability value is required, the "best convention to adopt is to double the observed one-tail probability [value]" [1476, p. 446].

In the discussion that followed, Jagger disagreed with the suggestion by Yates of doubling the one-tail probability value [678], as did Healy [604]. Cormack supported the doubling rule, but with reservations, stating that "support for the doubling rule must be *faute de mieux*"[14] [279, p. 455]; see also a 1986 article by Cormack on this topic [280]. Mantel also disagreed strongly with the recommendation by Yates of doubling the one-tail exact probability value to obtain a two-tailed exact test [884]; however, in a notable contretemps Mantel was to change his mind in 1990 and embraced the doubling rule [885]. Finally, Plackett observed that Fisher's argument for defining a two-tail probability value as twice the one-tail probability value was based on the practice of using nominal significance levels, now considered defective [1139, p. 458]. An alternative definition, Plackett noted, was introduced by Jerzy Neyman and Egon Pearson who arranged events in order of decreasing probability and calculated the total probability in the tail.

5.5 Mehta–Patel and a Network Algorithm

In 1983 Cyrus Mehta and Nitin Patel created an innovative network algorithm for the Fisher–Yates exact probability test for $r \times c$ contingency tables. The Mehta–Patel network algorithm eliminated the need to completely enumerate all possible contingency tables in the appropriate reference set.

C.R. Mehta

Cyrus Rustam Mehta earned his B.Tech. degree in civil engineering from the Indian Institute of Technology in 1967, his S.M. degree in management science from the Massachusetts Institute of Technology in 1970, and his Ph.D. in operations research from the Massachusetts Institute of Technology in 1973. In 1973 he was appointed Assistant Professor at the University of Pittsburgh. In 1977 Mehta left the University of Pittsburgh, accepting a Postdoctoral Fellowship with the Dana–Farber Cancer Institute in Boston, Massachusetts. In 1979 he joined the faculty at Harvard University as an Assistant Professor of Biostatistics. In 1984 he was promoted to Associate Professor and in 2000 he was promoted to Professor. In 1987, together with Nitin R. Patel, Mehta founded Cytel Software Corporation, where he is currently President.

[14] Loosely, "for lack of an alternative."

N.R. Patel

Nitin Ratilal Patel is a recognized expert on the development of fast and accurate computer algorithms to implement computationally-intensive statistical methods. Patel earned his Ph.D. in operations research from the Massachusetts Institute of Technology in 1973. He has been a visiting professor at the Massachusetts Institute of Technology since 1995. Previously, he was CMC Chair Professor at the Indian Institute of Management and has held visiting positions at Harvard University, the University of Michigan, the University of Montreal, and the University of Pittsburgh. In 1987 he co-founded Cytel Software Corporation with Cyrus Mehta, where he is presently Chairman and Chief Technology Officer. At Cytel, Patel played a leading role in the development of StatXact and LogXact, widely used software for exact non-parametric inference.

In 1983 Mehta and Patel developed a network algorithm for the Fisher–Yates exact probability test for unordered $r \times c$ contingency tables [920]. Unlike earlier algorithms by Freeman and Halton [480], March [890], and Baker [55] that were based on an exhaustive enumeration of all possible $r \times c$ contingency tables with fixed marginal frequency totals, the network algorithm proposed by Mehta and Patel circumvented the need to explicitly enumerate all the tables in the appropriate reference set. Earlier, in 1980, Mehta and Patel had presented a network algorithm for the exact treatment of $2 \times c$ contingency tables [919]. This 1983 paper extended the network algorithm presented in the 1980 paper to $r \times c$ contingency tables and was followed by a series of papers detailing the network algorithm and providing applications; see for example, a paper by Mehta, Patel, and Gray in 1985 [923] and two papers by Mehta and Patel in 1986 [921, 922].

A Network Algorithm

A network algorithm combines the best of combinatorics and graph theory. Originally, the purpose of a network algorithm was to select a path in a network along which to send network traffic (packets), such as a telephone network or a transportation network. Routing directs packets forward from a source to the ultimate destination, optimizing speed by selecting the shortest path through a series of intermediate nodes in the network. A network algorithm in a statistical context is similar, yet different.

Consider a reference set for an observed $r \times c$ contingency table comprised of all $r \times c$ contingency tables with the observed row and column marginal frequency totals. A network algorithm is a directed non-cyclic network consisting of nodes in a sequence of stages, corresponding to the reference

(continued)

set of $r \times c$ contingency tables. Distances between the nodes, called arcs, are defined so that the total distance of a path through the network corresponds to the value of the test statistic. At each intermediary node, the network algorithm computes the longest and shortest path for all paths passing through that node. The value of the test statistic is compared with the longest and shortest paths to determine (1) if all paths through the node contribute to the probability value, (2) if none of the paths through the node contributes to the probability value, or (3) if neither of these situations occurs.

If all paths through the node contribute to the probability value, the probability value is incremented and these paths are eliminated from further consideration. If none of the paths contributes to the probability value, they are also eliminated. Otherwise, the network algorithm continues and is concluded when all nodes have either been accounted for or have been eliminated [1342].

Following the notation of Mehta and Patel [921, 922], consider an $r \times c$ contingency table, $\mathbf{X}$, with non-negative entries x_{ij}, let R_i and C_j denote the row and column marginal frequency totals, respectively, for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, and define the reference set of all possible $r \times c$ contingency tables to be $\Omega\{\mathbf{Y}\}$, where $\mathbf{Y}$ is $r \times c$, $\sum_{j=1}^{c} y_{ij} = R_i$, and $\sum_{i=1}^{r} y_{ij} = C_j$. Under the null hypothesis of row and column independence, the hypergeometric probability of any $\mathbf{Y} \in \Omega$ can be expressed as a product of multinomial coefficients

$$p(\mathbf{Y}) = D^{-1} \prod_{j=1}^{c} \frac{C_j\,!}{y_{1j}\,!, \ldots, y_{rj}!},$$

where

$$D = \frac{N!}{R_1!, \ldots, R_r!}$$

and $N = \sum_{i=1}^{r} R_i = \sum_{j=1}^{c} C_j$. The probability value based on the Fisher–Yates exact test is defined as the sum of the hypergeometric probabilities for all contingency tables in Ω that are no more likely than the observed $r \times c$ contingency table $\mathbf{X}$. Specifically,

$$P = \sum_{\mathbf{Y} \in \omega} p(\mathbf{Y}),$$

where $\omega = \{\mathbf{Y}\colon \mathbf{Y} \in \Omega \text{ and } p(\mathbf{Y}) \leq p(\mathbf{X})\}$ [920].

As Mehta and Patel described the network algorithm, the set Ω can be represented as a network of nodes and arcs, wherein each path through the network represents one and only one contingency table $\mathbf{Y} \in \Omega$ and the length of the path

Fig. 5.2 An example of an observed 3 × 3 contingency table for network analysis

0	1	0	1
1	2	0	3
2	2	2	6
3	5	2	10

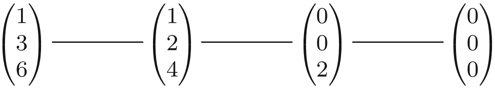


Fig. 5.3 Nodes and arcs of a single path for the observed data in Fig. 5.2

is $Dh(\mathbf{Y})$. The problem then is to identify and sum the lengths of all paths that are no longer than $Dh(\mathbf{X})$. The network is a directed graph with universal source S representing an empty table and universal sink U representing any filled table. The graph contains no cycles. All paths start in sink S, end in sink T, and have the same length, i.e., number of columns. The network is constructed in $c + 1$ stages, labeled successively $c, c - 1, \ldots, 0$ [921]. At any stage there exists a set of nodes, each labeled by a unique vector (k, R_k), where $R_k \equiv (R_{1k}, R_{2k}, \ldots, R_{rk})$. Arcs emanate from each node at stage k and every arc is directed to exactly one node at stage $k - 1$. The network is defined recursively by specifying all the nodes of the form $(k - 1, R_{k-1})$ that succeed node (k, R_k) and are connected to it by arcs. There is only one node at stage c, the initial or starting node S, which is labeled (c, R_c), where $R_c \equiv (R_1, R_2, \ldots, R_r)$. The result of the recursion is exactly one node at stage 0, the terminal node T.

An example analysis, although abbreviated, will clarify the process. Consider the observed 3 × 3 contingency table in Fig. 5.2 [cf. 1397]. The nodes and connecting arcs for a single path are successively displayed in Fig. 5.3 for the observed 3 × 3 contingency table in Fig. 5.2, where the first node $\{1\ 3\ 6\}$ represents the marginal frequency row totals (R_1, R_2, R_3) of the 3 × 3 contingency table in Fig. 5.2; the second node $\{1\ 2\ 4\}$ represents the marginal frequency row totals of the table in Fig. 5.2 with the x_{i1}, $i = 1, \ldots, 3$, values in column 1 $\{0\ 1\ 2\}$ removed; the third node $\{0\ 0\ 2\}$ represents the marginal frequency row totals of the table in Fig. 5.2 with the x_{i1} and x_{i2}, $i = 1, \ldots, 3$, values in columns 1 $\{0\ 1\ 2\}$ and 2 $\{1\ 2\ 2\}$ removed; and the last node containing $\{0\ 0\ 0\}$ represents an empty table with the x_{i1}, x_{i2}, and x_{i3}, $i = 1, \ldots, 3$, values in columns 1, 2, and 3 removed, i.e., $\{0\ 1\ 2\}$, $\{1\ 2\ 2\}$, and $\{0\ 0\ 2\}$.

The length of the first arc, the horizontal line between the first and second nodes, is given by the multinomial coefficient, $C_1!/(x_{11}!\ x_{21}!\ x_{31}!)$, based on the values of the difference between the first and second nodes,

$$\begin{pmatrix}1\\3\\6\end{pmatrix} - \begin{pmatrix}1\\2\\4\end{pmatrix} = \begin{pmatrix}0\\1\\2\end{pmatrix} \Longrightarrow \frac{3!}{0!\,1!\,2!} = 3\ ;$$

the length of the second arc between the second and third nodes is given by the multinomial coefficient, $C_2!/(x_{12}!\, x_{22}!\, x_{32}!)$, based on the values of the difference between the second and third nodes,

$$\begin{pmatrix}1\\2\\4\end{pmatrix} - \begin{pmatrix}0\\0\\2\end{pmatrix} = \begin{pmatrix}1\\2\\2\end{pmatrix} \Longrightarrow \frac{5!}{1!\,2!\,2!} = 30\ ;$$

and the length of the third arc between the third and last nodes is given by the multinomial coefficient, $C_3!/(x_{13}!\, x_{23}!\, x_{33}!)$, based on the values of the difference between the third and last nodes,

$$\begin{pmatrix}0\\0\\2\end{pmatrix} - \begin{pmatrix}0\\0\\0\end{pmatrix} = \begin{pmatrix}0\\0\\2\end{pmatrix} \Longrightarrow \frac{2!}{0!\,0!\,2!} = 1.$$

The length of the path is the product of the three multinomial values,

$$Dh(\mathbf{X}) = \prod_{j=1}^{c} \frac{C_j\,!}{x_{1j}\,!\, x_{2j}\,!\, x_{3j}\,!} = 3 \times 30 \times 1 = 90.$$

Then, the probability of the observed path $p(\mathbf{X})$ is the length of the path divided by the multinomial probability value based on the row marginal frequency totals of the full 3 × 3 contingency table in Fig. 5.2:

$$\frac{N\,!}{R_1!\, R_2!\, R_3!} = \frac{10!}{1!\,3!\,6!} = 840.$$

Thus, $90/840 = 0.1071$ is the Fisher exact hypergeometric point-probability value of the observed 3 × 3 contingency table in Fig. 5.2. This can be confirmed by calculating the exact point-probability value of the 3 × 3 contingency table in Fig. 5.2:

$$P(\mathbf{X}) = \frac{\displaystyle\prod_{i=1}^{r} R_i\,! \prod_{j=1}^{c} C_j\,!}{\displaystyle N\,! \prod_{i=1}^{r} \prod_{j=1}^{c} x_{ij}!} = \frac{1!\,3!\,6!\,3!\,5!\,2!}{10!\,0!\,1!\,0!\,1!\,2!\,0!\,2!\,2!\,2!} = \frac{54}{504} = 0.1071.$$

The full network diagram is depicted in Fig. 5.4 and contains a total of 14 nodes and 24 paths with each path corresponding to one of the possible 3 × 3 contingency tables in the reference set $\Omega\{\mathbf{Y}\}$, given the fixed row and column marginal frequency totals of $\{R_1, R_2, R_3\} = \{1, 3, 6\}$ and $\{C_1, C_2, C_3\} = \{3, 5, 2\}$, respectively. The exact hypergeometric probability of the observed table, denoted by the broken line

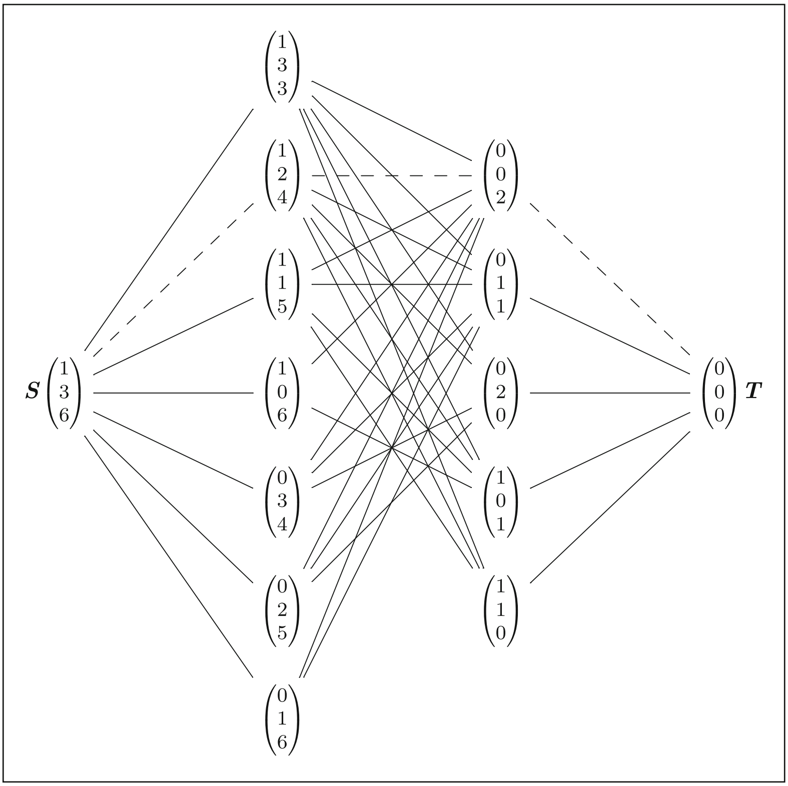


Fig. 5.4 Representation of the isomarginal family from Fig. 5.2 as an enumeration network. Each of the 24 contingency tables is represented by a path, denoted by a solid line, from node ***S*** to node ***T***, and each of the 14 nodes represents a sum of columns. The path for the observed contingency table is denoted by a *broken line*

in Fig. 5.4, is the sum of all the paths with lengths less than or equal to 90, divided by 840; in this case there are 23 paths with lengths less than or equal to $Dh(\mathbf{X}) = 90$ and the exact two-tailed probability value is the sum of the 23 lengths that are less than or equal to $Dh(\mathbf{X}) = 90$, divided by 840, yielding 0.8571. At the heart of the logic of the network algorithm is the stage-wise simultaneous processing of all stage c nodes before proceeding to stage $c - 1$.[15] If the path length is greater than $Dh(\mathbf{X})$ it cannot contribute to the probability value and that path is dropped from further consideration. Thus, explicit enumeration of all paths is unnecessary and, consequently, the network algorithm of Mehta and Patel is extremely efficient.

In 1988 Joe published an improvement to the network algorithm of Mehta and Patel [920] for computing the probability value of the Fisher–Yates exact test for

[15] Hirji and Johnson showed in 1996 that stage-wise processing is a very memory-intensive approach [630, p. 420].

unordered $r \times c$ contingency tables [688]. Mehta and Patel relied on upper and lower bounds, $Q_{\max}$ and $Q_{\min}$, respectively, to determine whether a path should be eliminated. Joe was able to refine the definitions of $Q_{\max}$ and $Q_{\min}$, resulting in more arcs being deleted earlier in the process. Joe found that the amount of computing time could be substantially decreased compared with the program of Mehta and Patel [921] and that the reduction was greatest when the column marginal frequency totals were quite different. In 1994, Valz and Thompson made further improvements to Mehta and Patel's original implementation of the network algorithm to further enhance its computational performance [1388].

In 1984 Saunders published FORTRAN subroutine ENUM to enumerate $r \times c$ contingency tables with repeated row totals [1223]. Saunders' subroutine demonstrated considerable savings in time when there was a duplication of row marginal frequency totals.

In 1985 Verbeek and Kroonenberg surveyed algorithms for testing independence in $r \times c$ contingency tables with fixed marginal frequency totals utilizing discrete methods [1397]. They noted that the three main approaches were complete enumeration, short-cuts avoiding enumeration of certain tables not in the critical region, and generation of a Monte Carlo sample from all possible tables with fixed marginal frequency totals, i.e., the isomarginal family.

Verbeek and Kroonenberg defined two basic structures of the available algorithms: (1) given a pair of marginal frequency distributions and a statistic S, find the distribution of S under the hypergeometric probability distribution of the contingency table, and (2) given a pair of marginal frequency distributions, a statistic S, and an observed value of S, S_{o}, find $p_{\text{o}} = p(S \geq S_{\text{o}})$ under the null distribution [1397, p. 162]. They then described a basic enumeration algorithm for (1) and three enumeration algorithms for (2); viz., a basic enumeration algorithm, a Monte Carlo algorithm, and a characteristic function algorithm. The rest of this lengthy article consists of a survey of existing algorithms for exact distributions of test statistics in $r \times c$ contingency tables with fixed marginal frequency totals and an extensive bibliography.

In 1985 Berry and Mielke developed two FORTRAN subroutines for an exact chi-squared test and the Fisher–Yates exact probability test, ECST (Exact Chi-Squared Test) and FEPT (Fisher Exact Probability Test), both for 2×2 contingency tables with fixed marginal frequency totals [127]. The procedures yielded the exact one- and two-tail probability values for each test. However, the Fisher–Yates exact probability test accumulated probability values for those 2×2 contingency tables with probability values equal to or less than the observed probability value, whereas the exact chi-squared test accumulated probability values for those 2×2 contingency tables with chi-squared values equal to or greater than the observed chi-squared value.

Given a 2×2 contingency table of n cases, with x denoting the observed frequency of any cell, and with r and c representing the row and column marginal frequency totals, respectively, of x, the minimum and maximum values of x are given by $v = \max(0, r + c - n)$ and $w = \min(r, c)$, respectively. The recursion over $v \leq x \leq w$ differed substantially from the recursion procedure described in the 1983 paper by Berry and Mielke [122] (q.v. page 284). The initial probability value for the recursion procedure in 1983 was obtained using Stirling's approximation for the required factorial expressions (q.v. page 227). In the 1985 paper the initial probability was assigned an arbitrary value q_v, a running total of the resulting q_i values was kept for $i = v, \ldots, w$, and at the end of the recursion each q_i value was divided by the total, yielding the desired exact probability values. This approach completely eliminated the calculation of the initial probability value (q.v. page 44).

In 1987 Berry and Mielke extended the capabilities of subroutines ECST and FEPT to 3 × 2 contingency tables with fixed marginal frequency totals [131], and in 1992 they presented high-speed recursion algorithms incorporating an arbitrary initial value for Fisher's exact probability test for 2 × 2, 2 × 3, 2 × 4, 2 × 5, 2 × 6, and 3 × 3 contingency tables with fixed marginal frequency totals, i.e., up to five degrees of freedom [949]. In 1994 Mielke, Berry, and Zelterman developed FORTRAN subroutine FEP222 based on a recursion algorithm with an arbitrary initial value for Fisher's exact probability test for 2 × 2 × 2 contingency tables with fixed marginal frequency totals [983], and in 1995 Zelterman, Chan, and Mielke provided algorithms to generate all possible 2×2×2 and 2×2×2×2 contingency tables with fixed marginal frequency totals [1489]. Finally, in 1996 Mielke and Berry presented FORTRAN subroutine EI222, based on a recursion algorithm with an arbitrary initial value for the first- and second-order interactions in 2×2×2 contingency tables with fixed marginal frequency totals [953].

In 1985 Berry and Mielke developed non-asymptotic permutation tests for Goodman and Kruskal's τ_a and τ_b statistics [534]. The algorithm was based on the exact mean, variance, and skewness under the conditional permutation distribution, which then employed the Pearson type III probability distribution to obtain approximate probability values [126]. They found the non-asymptotic approach to be superior to the conventional asymptotic method for small samples and for unbalanced marginal frequency distributions. The 1985 article was followed by a 1986 article that provided FORTRAN subroutine TAU for calculating the Goodman and Kruskal coefficients τ_b (row variable dependent), τ_a (column variable dependent), and the non-asymptotic probability value of coefficients as or more extreme than the observed values [130].

In 1985 Thakur, Berry, and Mielke provided a FORTRAN program for testing linear trend and homogeneity in proportions [1349]. Trend was evaluated by the Cochran–Armitage method as well as by multiple pairwise comparisons by the Fisher–Yates exact probability method. Again, a recursion technique with an arbitrary initial value was employed, yielding exact two-tailed probability values based on all permutations of cell frequencies with fixed marginal frequency totals.

In 1985 Mielke and Berry utilized a non-asymptotic permutation procedure based on the chi-squared statistic for an analysis of $r \times c$ contingency tables [947]. The method provided improved analyses for those cases where the conventional asymptotic approach was questionable, e.g., small expected cell frequencies. The non-asymptotic permutation approach was based on an algorithm that obtained the exact mean, variance, and skewness under the conditional permutation distribution, which was then utilized for inferences in conjunction with the Pearson type III probability distribution.[16] In 1986 Berry and Mielke developed FORTRAN subroutine CHI2 for analyzing $r \times c$ contingency tables with either highly disproportionate marginal frequency totals or relatively small marginal frequency totals, resulting in small expected cell frequencies [129].

In 1987 Zar presented a fast and efficient algorithm for the Fisher–Yates exact probability test for 2×2 contingency tables with fixed marginal frequency totals [1486]. The resulting program was written in BASIC and employed an initial probability value obtained with Stirling's factorial approximation (q.v. page 227) with subsequent probability values generated by recursion, as suggested previously by Feldman and Klinger [424] (q.v. page 220).

In 1988 Baglivo, Olivier, and Pagano proposed a new hybrid method for the analysis of $r \times c$ contingency tables with large and small cell frequencies [45]. The hybrid method consisted of two parts. The first part was done exactly and the remaining part was done with a normal density approximation. Given an $R \times C$ contingency table with observed cell frequencies x_{ij}, row marginal frequency totals r_i, column marginal frequency totals c_j and table total N, they partitioned the hypergeometric probability function for the entire $R \times C$ contingency table,

$$P(X) = \frac{\displaystyle\prod_{i=1}^{R} r_i! \prod_{j=1}^{C} c_j!}{\displaystyle N! \prod_{i=1}^{R} \prod_{j=1}^{C} x_{ij}!} ,$$

into products of conditional hypergeometric functions for 2×2 contingency tables as described by Pagano and Taylor Halvorsen in 1981 [1081] and Plackett in 1981 [1138] using a recursion method to subdivide $P(X)$. After factoring $P(X)$ as a product of two probability functions, they approximated the second of the two probability functions by a normal density function. Algorithms were included for the likelihood-ratio test, the Fisher–Yates exact probability test, and Pearson's chi-squared test of independence. In addition, Baglivo et al. extended the procedure to multidimensional contingency tables and illustrated the procedure with log-linear models.

Also in 1988, Balmer provided an algorithm for the recursive enumeration of $r \times c$ contingency tables with fixed row and column marginal frequency totals [56]. Simultaneously, the algorithm calculated the conditional probabilities given by row and column marginal frequency totals, according to both a hypergeometric and a multinomial model, evaluating the likelihood for the multinomial model across the complete family of tables [56, p. 290].

[16] See in this regard, a 1994 paper by Kulinskaya on "Large sample results for permutation tests of association" published in *Communications in Statistics—Theory and Methods* [780].

In 1988 Berry and Mielke revisited the non-asymptotic chi-squared test that they had previously developed in 1985 [134] (q.v. page 293). They employed extensive Monte Carlo procedures to compare asymptotic chi-squared and likelihood-ratio tests with the non-asymptotic chi-squared test for sparse $r \times c$ contingency tables. Monte Carlo comparisons of five contingency table tests under the null hypothesis were considered:

1. A non-asymptotic chi-squared test.
2. An asymptotic chi-squared test with a normal distribution.
3. An asymptotic chi-squared test with a chi-squared distribution.
4. An asymptotic likelihood-ratio test with a normal distribution.
5. An asymptotic likelihood-ratio test with a chi-squared distribution.

They made a total of 270 comparisons involving the five tests, three contingency table sizes (2×2, 2×4, and 3×4), three sample sizes (20, 40, and 80), three marginal configurations (one equal and two unequal), and two models (independence and homogeneity). A total of 540,000 contingency tables were examined. All probability values were based on 10,000 randomly generated contingency tables [134].

Berry and Mielke concluded that, for sparse contingency tables with small degrees of freedom, the asymptotic chi-squared test with a normal distribution and the likelihood-ratio test with a normal distribution had "no utility," the chi-squared test with a chi-squared distribution and the likelihood-ratio test with a chi-squared distribution were "not very satisfactory" and "poor," respectively, while the non-asymptotic chi-squared test was clearly superior to the other four tests for sparse $r \times c$ contingency tables [134, p. 259].

In 1990 Agresti, Mehta, and Patel extended the network algorithm of Mehta and Patel [920] to $r \times c$ contingency tables with ordered categories [6]. In 1991 Vollset and Hirji presented a microcomputer program for calculating exact and asymptotic tests and confidence intervals for a binomial proportion and the common odds-ratio for both a single and a series of 2 × 2 contingency tables [1399]. The program was written in GAUSS and the authors demonstrated that it was considerably faster than previously published programs, including the Mehta and Patel algorithm contained in StatXact.

Also in 1991, Vollset, Hirji, and Elashoff continued the work of Vollset and Hirji [1399], proposing three modifications of the network algorithm of Mehta, Patel, and Gray [923] that enhanced computational efficiency exceeding an order of magnitude [1400]. They also compared the modified method with the fast Fourier transform algorithm of Pagano and Tritchler (q.v. page 338) [1082, 1083], noting that the fast Fourier transform algorithm was not as efficient or reliable as the network algorithm, and concluded that "we will not recommend its use" [1400, p. 408].

In 1992 Mielke and Berry provided algorithms and associated FORTRAN subroutines for the Fisher exact probability test for $r \times c$ contingency tables up to five degrees of freedom: 2 × 2, 2 × 3, 2 × 4, 2 × 5, 2 × 6, 3 × 3 and 2 × 2 × 2 [949]. The same general algorithm was used for each of the seven subroutines, which was based on a recursion technique developed by Adolphe Quetelet in 1846 for calculating binomial probability values [1147, p. 260].

L.A.J. Quetelet

Lambert Adolphe Jacques Quetelet, born 22 February 1795 in Ghent, was a Belgium astronomer and mathematician, one of the founders of the Royal Statistical Society, tutor to Queen Victoria's husband, Prince Albert [738, p. 198], and a forerunner in demonstrating the importance of statistics to social science [289]. Among those who influenced Quetelet were Thomas Malthus, Joseph Fourier, and Pierre-Simon Laplace [805, p. 280]. Quetelet is considered the father of quantitative social science with his concepts of "l'homme moyen" (the average man) introduced in his 1835 essay *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale* or "On Man and the Development of his Faculties, or Essays on Social Physics" in which he described his concept of the average man characterized by the mean values of measured variables that followed a normal distribution [289,428,1321, pp. 51–65]. Quetelet was convinced that knowledge of causes influenced the course of human affairs, rather than the generally accepted "hand of God" in the early 1800s.

In "Essays on Social Physics" Quetelet calculated the average height and weight of subjects and cross-tabulated these with sex, age, occupation, and geographical region. In combination, these average values produced what Quetelet called "the average man," believing that if the average man could be ascertained for one nation, he could represent that nation [289]. Quetelet felt that no individual was free from the "laws" that governed him. By laws Quetelet meant counting repetitions of a frequently occurring social act, such as, for example, suicide. Florence Nightingale identified the laws of Quetelet as an answer to the amelioration of social life and in her eulogy she considered Quetelet as "the founder of the most important science in the whole world" [912, p. 190]. Lambert Adolphe Jacques Quetelet died just 5 days short of his 79th birthday on 17 February 1874 in Brussels.

Beginning with an arbitrarily chosen initial value (in this case, 1), a recursion procedure generated relative frequency values for all possible contingency tables with the observed row and column marginal frequency totals (q.v. page 44). The required probability value was obtained by summing the relative frequency values equal to or less than the observed relative frequency value and dividing by the unrestricted relative frequency total. Consequently, no factorial expressions, logarithms, or log-factorial values were required.

In 1993 Baglivo, Olivier, and Pagano published a lengthy article on the analysis of discrete data in which they advocated rerandomization (resampling) permutation methods for a number of models and tests, including multinomial testing and goodness of fit of log-linear models for contingency tables [47]. They presented algorithms that were different from other proposed methods in that they showed how to calculate the permutation distributions of commonly-used statistics, rather than simply calculating probability values for exact tests [47, p. 175].

In 1996 Hirji and Johnson compared the speed and accuracy of the network algorithm of Mehta and Patel [920] with the fast Fourier transform algorithm of Baglivo, Olivier, and Pagano [46] for unordered $2 \times c$ contingency tables, showing that the two algorithms rest on the same foundation: a recursive polynomial relation [630]. However, the network algorithm is equivalent to a stage-wise implementation of the recursion [630, p. 424], while the fast Fourier transform algorithm is based on performing the recursion at complex roots of unity [630, pp. 424–425]. In an examination of three $2{\times}3$, four $2{\times}6$, one $2{\times}7$, and four $2{\times}9$ contingency tables with varying marginal frequency totals, they showed that for the Pearson chi-squared, likelihood-ratio, and Freeman–Halton statistics the network algorithm of Mehta and Patel [920] was more efficient and accurate than the fast Fourier transform algorithm of Baglivo, Olivier, and Pagano [46].

In 1997 Shao proposed an efficient algorithm for computing the Fisher–Yates exact probability test for unordered $2 \times c$ contingency tables [1255]. When all or some of the column marginal frequency totals were identical, the improved algorithm substantially reduced the computational effort needed to obtain an exact probability value. Shao noted that his new algorithm was applicable to exact Pearson chi-squared and exact likelihood-ratio tests of independence and Shao provided several numerical examples with which to compare the computing time of the improved algorithm and the original network algorithm of Mehta and Patel [920].

5.5.1 Multi-Way Contingency Tables

Increasingly in this period, interest developed in analyzing r-way contingency tables. In 1988 Mielke and Berry [948] developed methods for analyzing independence in r-way contingency tables based on the exact first three cumulants of both the classical Pearson chi-squared statistic and a modification of the chi-squared statistic by Zelterman [1488]. Mielke and Berry presented methods to analyze $n_1 \times \cdots \times n_r$ contingency tables and goodness-of-fit frequency data based on the hypergeometric probability distribution conditioned on fixed marginal frequency totals.

Consider an r-way contingency table consisting of $n_1 \times n_2 \times \cdots \times n_r$ cells where the observed frequency of the $(j_1, \ldots, j_r)$th cell is denoted by $o_{j_1,\ldots,j_r}$, the marginal frequency total associated with subscript j_i of category i is denoted by

$$\langle i \rangle_{j_i} = \sum_{*|j_i} o_{j_1,\ldots,j_r}$$

for $j_i = 1, \ldots, n_i$, $i = 1, \ldots, r$, and $\sum_{*|j_i}$ is the partial sum over all cells with subscript j_i fixed. Then, the frequency total of the entire r-way contingency table is given by

$$N = \sum_{j_i=1}^{n_i} \langle i \rangle_{j_i}$$

for $i = 1, \ldots, r$. If $p_{j_1,\ldots,j_r} \geq 0$ is the probability that any of the N total events occurs in the $(j_i, \ldots, j_r)$th cell, then the multinomial probability is given by

$$P\left(o_{j_1,\ldots,j_r}\right) = \left(N! \Bigg/ \prod_{i=1}^{r} \prod_{j_i=1}^{n_i} o_{j_1,\ldots,j_r}!\right) \left(\prod_{i=1}^{r} \prod_{j_i=1}^{n_i} p_{j_1,\ldots,j_r}{}^{o_{j_1,\ldots,jr}}\right),$$

where $0^0 = 1$. The assumed positive marginal probability associated with subscript j_i is given by

$$[\,i\,]_{j_i} = \sum_{*|j_i} p_{j_1,\ldots,j_r}$$

for $j_i = 1, \ldots, n_i$, $i = 1, \ldots, r$, and

$$\sum_{j_i=1}^{n_i} [\,i\,]_{j_i} = 1$$

for $i = 1, \ldots, r$. Then, the marginal multinomial probability associated with category i is given by

$$P\left(\langle i \rangle_{j_i}\right) = \left(N! \Bigg/ \prod_{j_i=1}^{n_i} \langle i \rangle_{j_i}!\right) \left(\prod_{j_i=1}^{n_i} [\,i\,]_{j_i}^{\langle i \rangle_{j_i}} \right)$$

for $i = 1, \ldots, r$. The null hypothesis that the r categories are independent specifies that

$$p_{j_1,\ldots,j_r} = \prod_{i=1}^{r} [\,i\,]_{j_i} > 0$$

and the conditional distribution function of the r-way contingency table under the null hypothesis, H_0, is given by

$$P\left(o_{j_1,\ldots,j_r} \mid \langle 1 \rangle_{j_1}, \ldots, \langle r \rangle_{j_r},\ H_0\right) = \frac{P\left(o_{j_1,\ldots,j_r} \mid H_0\right)}{\displaystyle\prod_{i=1}^{r} P\left(\langle i \rangle_{j_i}\right)}.$$

Algebraic manipulation then yields the hypergeometric distribution function given by

$$P\left(o_{j_1,\ldots,j_r} \mid \langle 1\rangle_{j_1},\ldots,\langle r\rangle_{j_r},\, H_0\right)=\frac{\displaystyle\prod_{i=1}^{r}\prod_{j_i=1}^{n_i}\langle i\rangle_{j_i}!}{\displaystyle(N!)^{r-1}\prod_{i=1}^{r}\prod_{j_i=1}^{n_i}o_{j_1,\ldots,j_r}!}\,,$$

which is independent of any unknown probabilities under the null hypothesis [948].[17] Thus, the marginal frequency totals, $\langle i\rangle_{j_i}$, are sufficient statistics for the marginal multinomial probabilities, $[\,i\,]_{j_i}$, under the null hypothesis. This hypergeometric distribution function provided the basis for testing the independence of categories for any r-way contingency table. In 1989 Berry and Mielke [136] released FORTRAN subroutine RWAY for testing independence in r-way contingency tables using the non-asymptotic moment-approximation described in Mielke and Berry [948], and in 1994 Berry and Mielke released FORTRAN subroutine GOF which used the methods of Mielke and Berry [948] to test for goodness of fit between observed category frequencies and the a priori category probabilities [140]. Subsequently, a resampling-approximation probability procedure for r-way contingency tables with fixed marginal frequency totals was developed by Mielke, Berry, and Johnston in 2007 [975].

5.5.2 Additional Contingency Table Analyses

In 1991 Cormack and Mantel investigated the Fisher–Yates exact probability test for 2×2 contingency tables [281]. Employing extensive computer simulations, they attempted to resolve the long-standing controversy as to whether the row and column marginal frequency distributions should be considered as both fixed, one fixed and the other random, or both random. They concluded that both marginal frequency distributions should be considered as fixed, lending credence to some 55 years of published work on permutation methods based on fixed marginal frequency totals.

As Cormack and Mantel noted, a second problem with the Fisher–Yates exact probability test was that of defining the second-tail probability value in asymmetric cases; see for example, articles by Cormack in 1986 [280] and Mantel in 1990 [885]. The second-tail problem, as it is known, is whether to double the tail-probability value of the tail in which the observed table configuration lies (doubling rule) or to sum the probability values in the second tail that are equal to or less than the probability value of the observed table and add that sum to the first-tail probability value (Irwin's rule). When the distribution is symmetric (i.e., all four marginal frequency totals are identical) both rules yield the same result. In this exploration of

[17] For a rigorous proof of the exact contingency formula, see a 1969 article by John Halton in *Mathematical Proceedings of the Cambridge Philosophical Society* [578].

the Fisher–Yates exact probability test, Cormack and Mantel "deliberately bypassed the problem of the second tail" by looking only at symmetric cases [281, p. 33].[18]

In 1992 Berry and Mielke introduced a new multivariate measure of association for a nominal independent variable and nominal, ordinal, or interval dependent variables [138]. The measure of association, A, was a chance-corrected multivariate measure and was applicable to a nominal independent variable and any combination of nominal, ordinal, and interval dependent variables. Because the dependent variables may possess different units of measurement, they must be made commensurate. Berry and Mielke chose Euclidean commensuration.

Commensuration

Often when variables possess different units of measurement, they need to be made commensurate, i.e., standardized to a common unit of measurement. For a simple example, consider two disparate variables: direction measured in radians and denoted by w_1, and speed measured in miles per hour and denoted by w_2. To make a proper comparison of the two variables, w_1 and w_2 must be made commensurate, otherwise the large units of w_2 (mph) would completely overwhelm the smaller units of w_1 (rads). Two types of commensuration are commonly used: Euclidean commensuration and Hotelling commensuration.

Euclidean Commensuration. Let $y'_I = [y_{1I}, \ldots, y_{rI}]$, $I = 1, \ldots, N$, denote N non-commensurate r-dimensional values for $r \geq 2$. The corresponding N Euclidean-commensurate r-dimensional values of $x'_I = [x_{1I}, \ldots, x_{rI}]$ for $I = 1, \ldots, N$ are given by $x_{jI} = y_{jI}/\phi_j$, where

$$\phi_j = \left[\sum_{I<J} \left| y_{jI} - y_{jJ} \right|^v \right]^{1/v} .$$

As defined, the Euclidean commensurated data have the property that

$$\sum_{I<J} \left| x_{jI} - x_{jJ} \right|^v = 1$$

for $j = 1, \ldots, r$ and any $v > 0$. Usually, Euclidean commensuration is associated with $v = 1$ [138]. This commensuration procedure is based on the distance between the r response measurements of subjects ω_I and ω_J and is given by the distance function

(continued)

[18] Nathan Mantel long held to Irwin's rule, e.g., [883, p. 379], but recanted in 1990 to support doubling the observed one-tail probability value [885, p. 369].

$$\Delta_{I,J} = \left[\sum_{j=1}^{r} \left(x_{jI} - x_{jJ} \right)^2 \right]^{v/2} ,$$

where $v > 0$.

Hotelling Commensuration. An alternative commensuration, termed Hotelling commensuration, is based on the distance function

$$\Delta_{I,J} = \left[(y_I - y_J)' \, S^{-1} \, (y_I - y_J) \right]^{v/2} ,$$

where S is the $r \times r$ variance-covariance matrix given by

$$S = \begin{bmatrix} \frac{1}{N} \sum_{I=1}^{N} (y_{1I} - \bar{y}_1)^2 & \cdots & \frac{1}{N} \sum_{I=1}^{N} (y_{1I} - \bar{y}_1)(y_{rI} - \bar{y}_r) \\ \vdots & & \vdots \\ \frac{1}{N} \sum_{I=1}^{N} (y_{rI} - \bar{y}_r)(y_{1I} - \bar{y}_1) & \cdots & \frac{1}{N} \sum_{I=1}^{N} (y_{rI} - \bar{y}_r)^2 \end{bmatrix} ,$$

$v > 0$, and

$$\bar{y}_j = \frac{1}{N} \sum_{I=1}^{N} y_{jI}$$

for $j = 1, \ldots, r$ [943,951]. Usually, Hotelling commensuration is associated with $v = 2$.

As Berry and Mielke noted, Euclidean commensuration ensured that the resulting inferences were independent of the units of the individual response measurements, and invariant to linear transformations of the response measurements [138, p. 43]. The exact probability value for A was the proportion of all possible values of A equal to or greater than the observed value of A. For larger samples, Berry and Mielke recommended a moment-approximation permutation procedure based on the first three exact moments of the Pearson type III distribution. Later, in 1992, Berry and Mielke provided FORTRAN subroutines EMAP (Exact Multivariate Association Procedure) and AMAP (Approximate Multivariate Association Procedure), which calculated the chance-corrected measure of association, A, and its associated exact and approximate probability values [139].

In 1993 Mielke and Berry presented algorithms and associated FORTRAN subroutines for exact goodness-of-fit probability tests [950]. The exact subroutines were conditional and utilized a recursion procedure with an arbitrary initial value to generate relative frequency values for all possible configurations of N objects in k categories. The required probability values were obtained by summing the relative frequency values equal to or less than the observed relative frequency value and dividing by the unrestricted frequency total (q.v. page 44).

5.6 MRPP and the Pearson Type III Distribution

As P.K. Sen noted in 1965, although permutation tests are quite easy to define and have many desirable properties, the main difficulty with all permutation methods is the labor of numerical computation involved in generating the permutation distribution of the statistic in question, especially with large samples. Thus, as Sen noted, some simple approximations to these discrete permutation distributions are more or less essential [1247, p. 106]. Consequently, the computation of exact permutation probability values early on was necessarily limited to small samples—often very small samples. For larger samples, researchers often relied on approximating the discrete permutation distribution with the beta distribution, typically utilizing two, three, or four moments; see for example, a 1938 article by Pitman [1131] (q.v. page 81). The popularity of the beta distribution was most likely due to its strong relationship to the distribution of Student's t statistic and to the distribution of Fisher's variance-ratio z statistic, and subsequently to the distribution of Snedecor's F statistic [1430, p. 353].

The use of the beta distribution required standardization of the permutation test statistic to ensure that the statistic varied between 0 and 1. For example, both Pitman in 1937 and Welch in 1937 defined

$$W = \frac{SS_{\text{Error}}}{SS_{\text{Error}} + SS_{\text{Treatment}}},$$

instead of t or z, as W was constructed to vary between 0 and 1 [1129, 1428].

In 1943 Scheffé had been sharply critical of the use of moments to approximate discrete permutation distributions, stating that in his opinion the justification of moment approximations was never satisfactory from a mathematical point of view [1230, p. 311]. Although Scheffé does not mention the beta distribution specifically, it was so widely used at that time that it can be assumed with some confidence that Scheffé included the beta distribution in his criticism. Moreover, some test statistics are not amenable to moment-approximations because of their intractable structure, e.g., double ratios [969].

In 1981 Mielke, Berry, and Brier replaced the beta distribution with the Pearson type III distribution[19] in applications of MRPP (q.v. page 254), due to the difficulty of making simple associations between the parameters of the beta distribution and the moments of the permutation distribution, even after reparameterization [934, 968]. The Pearson type III distribution, as a three-parameter gamma distribution, had the advantage of being totally characterized by the exact mean, variance, and skewness, in the same manner that the normal distribution, as a two-parameter distribution, is fully characterized by the exact mean and variance—a property that does not hold for the beta distribution.[20] An added advantage of the Pearson type III distribution is that when the skewness parameter is zero, the distribution is normal. In describing the Pearson type III distribution, Pearson noted "[t]his generalized probability curve fits with a high degree of accuracy a number of measurements and observations hitherto not reduced to theoretical treatment" [1104, p. 331]. An SPSS implementation of MRPP utilizing the Pearson type III distribution was later published by Cai in 2006 [236] and a SAS/IML procedure for MRPP was published somewhat earlier by Johnson and Mercante in 1993 [694].

The Pearson type III approximation depends on the exact mean, variance, and skewness of δ (q.v. page 255) under the null hypothesis given by

$$\mu_\delta = \frac{1}{M}\sum_{I=1}^{M}\delta_I\,,$$

$$\sigma_\delta^2 = \frac{1}{M}\sum_{I=1}^{M}(\delta_I - \mu_\delta)^2\,,$$

and

$$\gamma_\delta = \left[\frac{1}{M}\sum_{I=1}^{M}(\delta_I - \mu_\delta)^3\right]\Big/\sigma_\delta^3\,,$$

respectively. In particular, the standardized statistic given by

$$T = \frac{\delta - \mu_\delta}{\sigma_\delta}$$

[19] The Pearson type III distribution was one of four distributions introduced by Karl Pearson in 1895 [1106], although the type III distribution had previously been presented without discussion by Pearson in 1893 [1104, p. 331]. The type V distribution introduced by Pearson in 1895 was simply the normal distribution and the Pearson type I distribution was a generalized beta distribution.

[20] Mielke, Berry, and Brier were, of course, not the first to use the Pearson type III distribution to approximate a discrete permutation distribution. B.L. Welch utilized the Pearson type III distribution in a paper on the specification of rules for rejecting too variable a product [1427, p. 47] and used it again in a paper on testing the significance of differences between the means of two independent samples when the population variances were unequal [1430, p. 352].

is presumed to follow the Pearson type III distribution with density function given by

$$f(y) = \frac{(-2/\gamma_\delta)^{4/\gamma_\delta^2}}{\Gamma\left(4/\gamma_\delta^2\right)} \left[-(2 + y\gamma_\delta)/\gamma_\delta\right]^{(4-\gamma_\delta^2)/\gamma_\delta^2} \exp(-2(2 + y\gamma_\delta)/\gamma_\delta^2)$$

when $-\infty < y < -2/\gamma_\delta$ and $\gamma_\delta < 0$, or

$$f(y) = \frac{(2/\gamma_\delta)^{4/\gamma_\delta^2}}{\Gamma\left(4/\gamma_\delta^2\right)} \left[(2 + y\gamma_\delta)/\gamma_\delta\right]^{(4-\gamma_\delta^2)/\gamma_\delta^2} \exp(-2(2 + y\gamma_\delta)/\gamma_\delta^2)$$

when $-2/\gamma_\delta < y < \infty$ and $\gamma_\delta > 0$, or

$$f(y) = (2\pi)^{-1/2} \exp(-y^2/2)$$

when $\gamma_\delta = 0$, i.e., the standard normal distribution. If δ_{o} and

$$T_{\mathrm{o}} = \frac{\delta_{\mathrm{o}} - \mu_\delta}{\sigma_\delta},$$

are the observed δ and T values, then

$$P(\delta \leq \delta_{\mathrm{o}} \,|\, H_0) \doteq \int_{-\infty}^{T_{\mathrm{o}}} f(y)\, dy$$

and

$$P(\delta \geq \delta_{\mathrm{o}} \,|\, H_0) \doteq \int_{T_{\mathrm{o}}}^{\infty} f(y)\, dy$$

denote approximate probability values, which are evaluated numerically over an appropriate finite interval. The Pearson type III distribution is used to approximate the permutation distribution of T because it is completely specified by γ_δ and includes the normal and chi-squared distributions as special cases. This approximation allows for the substantial negative skewness often encountered under the null hypothesis [935, 936]. Thus, these distributions are asymptotic limits of the permutation distribution for some situations.

5.7 MRPP and Commensuration

Whenever there are two or more responses for each object, the response measurements may be expressed in different units. It is then necessary to standardize the response variables to a common unit of measurement prior to statistical analysis; a process termed commensuration (q.v. page 301). In 1991 Mielke established

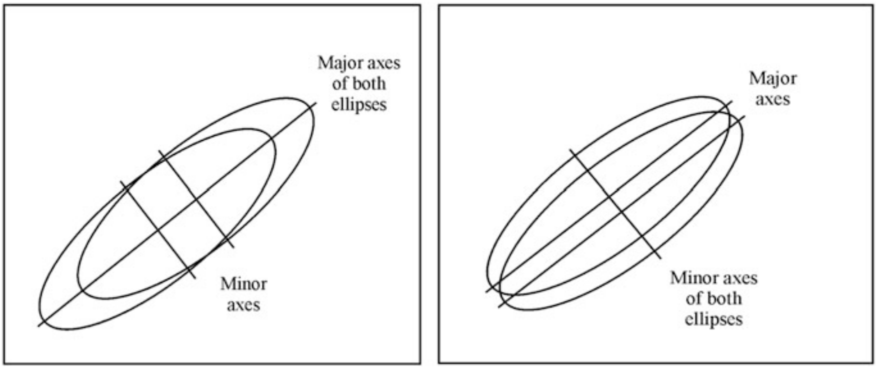


Fig. 5.5 A bivariate location shift parallel to the major axes, on the left, favoring Euclidean commensuration, and a bivariate location shift parallel to the minor axes, on the right, favoring Hotelling commensuration

that with $v = 2$, both the Hotelling T^2 [652] and the Bartlett–Nanda–Pillai trace statistics [79, 1020, 1128] are but special cases of MRPP with Hotelling commensuration [943, 965, Sect. 2.10, pp. 53–57]. In 1994 Mielke and Berry showed that, for multivariate tests, both Euclidean commensuration and Hotelling commensuration were far more robust with $v = 1$ than with $v = 2$ [951] (q.v. page 400).

In 1999 Mielke and Berry demonstrated that when analyzing correlated bivariate data, Euclidean commensuration yielded more powerful tests than Hotelling commensuration when the bivariate location shift was parallel to the major axes [958]. On the other hand, Hotelling commensuration yielded a more powerful test than Euclidean commensuration when the bivariate location shift was parallel to the minor axes. Figure 5.5 illustrates a bivariate location shift that is parallel to the major axes on the left and a bivariate location shift that is parallel to the minor axes on the right; see also a 1999 article by Pellicane and Mielke in *Wood Science and Technology* [1115].

5.8 Tukey and Rerandomization

In June of 1988 John Tukey (q.v. page 232) read a paper at the Ciminera Symposium in honor of Joseph L. Ciminera held in Philadelphia, Pennsylvania. The paper was titled "Randomization and rerandomization: The wave of the past in the future" [1382]. Although Tukey mentioned that the paper had been submitted for publication, it apparently was never published. However, copies of this important paper have survived.[21]

[21] Authors' note: special thanks to Charles Greifenstein, Manuscript Librarian at the Library of the American Philosophical Society in Philadelphia, for retrieving this manuscript from their extensive

Tukey began the paper by redefining the “three R’s” as Randomization, Robustness, and Rerandomization. When Tukey wrote “randomization,” he meant a controlled randomized design wherein treatments were randomly assigned to subjects in an effort to eliminate bias and to nearly balance whatever is important [1382, p. 17]. When Tukey wrote “robustness,” he meant to ensure high stringency, high efficiency, and high power over a wide range of probability models [1382, p. 17]. And when Tukey wrote “rerandomization,” he meant analysis of randomized comparative experiments by means of permutation methods to confine the probabilities to those we have ourselves made [1382, p. 17].

Tukey distinguished among three types of rerandomization. First, complete rerandomization, i.e., exact permutation analysis; second, sampled rerandomization, i.e., resampling permutation analysis; and third, subset rerandomization, e.g., double permutation analysis. Long an advocate of permutation methods [216], it is in this paper that Tukey refers to rerandomization as the “[p]latinum [s]tandard” of significance tests. After critically denouncing techniques such as the bootstrap and the jackknife, Tukey concluded the paper by arguing that when an experiment can be randomized, it should be. Then, the preferred analysis method should be based on rerandomization. In an important affirmation of permutation methods, he stated that “[n]o other class of approach provides significance information of comparable quality” [1382, p. 18]. This is consistent with a statement by Efron and Tibshirani, promoters of the bootstrap, writing in 1993: “[w]hen there *is* something to permute . . . it is a good idea to do so, even if other methods like the bootstrap are brought to bear” [402, p. 218].[22]

The Jackknife

The jackknife (and the bootstrap) are often considered as alternatives to permutation procedures (q.v. page 8). It is interesting that Tukey decried the use of the jackknife as he is often given credit for promoting its use and for providing the term “jackknife” to identify the procedure.

The jackknife procedure is a cross-validation technique first developed by Maurice Quenouille to estimate the bias of an estimator [1145, 1146]. John Tukey expanded the use of the jackknife to include variance estimation and coined the term “jackknife” because like a jack-knife—such as a Swiss Army knife or a Boy Scout pocket knife—this technique has wide applicability to many different problems, but is inferior for those problems for which special tools have been designed [1, 211]. The idea underlying the jackknife is simply to divide a sample of observations into many subsamples and

(continued)

collection of papers, articles, and books by John W. Tukey, which was donated to the American Philosophical Society by the estate of John Wilder Tukey in 2002.

[22] Emphasis in the original.

compute a statistic of interest for each subsample. Each statistic then provides information about the distribution of the parameter of interest [1176, p. 82]. The most popular of the jackknife procedures is the "drop-one jackknife," wherein n subsamples of size $n - 1$ are generated and examined.

Tukey is most often cited as providing the original suggestion for variance estimation, but the citation is to a brief piece of only seven sentences published in *The Annals of Mathematical Statistics* in 1958 consisting primarily of a report presented at a 3–5 April 1958 meeting of the Institute of Mathematical Statistics in Ames, Iowa. Robinson and Hamann [1176, p. 84] reported that the earliest explicit reference to the term "jackknife" in a peer-reviewed publication was either David Brillinger [211] or Rupert Miller [994] in 1964. H.A. David gave credit to Miller as being the first [323]. It should be noted that Brillinger attributed the first use of the term jackknife to Tukey in a 1959 unpublished manuscript [1377], while Miller attributes first use to Tukey in a 1962 unpublished manuscript [1380]. For the origins of the jackknife procedure see especially discussions by Robinson and Hamann [1176, pp. 83–84], Miller [994, 995], Brillinger [211], and Abdi and Williams [1].

5.9 Matched-Pairs Permutation Analysis

In 1982 Mielke and Berry again utilized the Pearson type III probability distribution in a presentation of a class of permutation methods for matched pairs based on distances between each pair of signed observed values [945]. Let $x_i = d_i z_i$ denote the usual matched-pairs model for $i = 1, \ldots, n$, where d_i is a fixed positive score, z_i is either $+1$ or -1, and the null hypothesis specifies that $P\{z_i = 1\} = P\{z_i = -1\} = 1/2$. Mielke and Berry considered the test statistic given by

$$\delta = \binom{n}{2}^{-1} \sum_{i<j} |x_i - x_j|^v ,$$

where $v > 0$ and the sum is over all $\binom{n}{2}$ combinations of the integers from 1 to n. Under the null hypothesis, the mean, variance, and skewness of δ are given by

$$\mu_\delta = \frac{A}{n(n-1)},$$

$$\sigma_\delta^2 = \frac{B}{[n(n-1)]^2},$$

and

$$\gamma_\delta = \frac{-6C}{B^{3/2}},$$

respectively, where

$$A = \sum_{i<j} \left(a_{ij} + b_{ij}\right),$$

$$B = \sum_{i<j} \left(a_{ij} - b_{ij}\right)^2,$$

$$C = \sum_{i<j<k} \left(a_{ij} - b_{ij}\right)\left(a_{ik} - b_{ik}\right)\left(a_{jk} - b_{jk}\right),$$

$a_{ij} = \left|d_i + d_j\right|^v$, $b_{ij} = \left|d_i - d_j\right|^v$, and the sums for A, B, and C are over all $\binom{n}{2}$ and $\binom{n}{3}$ combinations of the integers from 1 to n, respectively. The critical regions of these tests correspond to small values of δ and the probability associated with a realized value of δ, δ_{o}, is approximated by

$$P\{\delta \leq \delta_{\text{o}}\} \doteq \int_{-\infty}^{T_{\text{o}}} f(u)du,$$

where

$$T_{\text{o}} = \frac{\delta_{\text{o}} - \mu_\delta}{\sigma_\delta}$$

and $f(u)$ is the density function of the Pearson type III distribution (q.v. page 305).

Mielke and Berry observed that the scores, d_i for $i = 1, \ldots, n$, associated with these tests may be the observed values of matched-pairs differences or transformations of the observed values such as power or rank transformations [945]. Thus, if $d_i = 1/2$ for $i = 1, \ldots, n$, then

$$\mu_\delta = \frac{1}{2},$$

$$\sigma_\delta^2 = \frac{1}{2n(n-1)},$$

$$\gamma_\delta = \frac{-4(n-2)}{[2n(n-1)]^{1/2}},$$

and the test of δ is equivalent to the two-sided version of the sign test.

If $v = 2$ and $d_i = r_i$ for $i = 1, \ldots, n$, where the r_i are the rank-order statistics from below and there are no ties, then

$$\mu_\delta = \frac{2\left(n+1\right)\left(n+\frac{1}{2}\right)}{3},$$

$$\sigma_\delta^2 = \frac{8\left(n+1\right)\left(n^2-\frac{1}{4}\right)\left(n+\frac{6}{5}\right)}{9n\left(n-1\right)},$$

$$\gamma_\delta = \frac{-4\left(n^2-\frac{7n}{2}+3\right)\left(n^2+\frac{13n}{5}+\frac{12}{7}\right)}{\left[2n\left(n^2-1\right)\left(n^2-\frac{1}{4}\right)\right]^{1/2}\left(n+\frac{6}{5}\right)^{3/2}},$$

and the test of δ is equivalent to the two-sided version of the Wilcoxon signed-ranks test. In addition, if R^+ and R^- are the absolute sums of the positive and negative signed ranks, respectively, then the identity relating δ, R^+, and R^- is given by

$$\delta = \frac{n\left(n+1\right)\left(2n+1\right)}{3\left(n-1\right)} - \frac{2\left(R^+-R^-\right)^2}{n\left(n-1\right)}.$$

Also, the matched-pairs t test statistic given by $t = n\bar{x}/s_x$, where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2,$$

is a special case of δ when $v = 2$. The identities specifying the association between δ and t are given by

$$\delta = \frac{2}{t^2+n-1}\sum_{i=1}^{n} x_i^2 \qquad \text{and} \qquad t^2 = \frac{2}{\delta}\sum_{i=1}^{n} x_i^2 - n + 1.$$

The remainder of the article was devoted to power comparisons of $v = 1$ and $v = 2$ based on $n = 20$ and $n = 80$ for five distributions: double exponential, logistic, normal, uniform, and a U-shaped distribution. Mielke and Berry concluded that δ statistics based on $v = 1$ showed statistical advantages over those based on $v = 2$, especially for a wide variety of heavy- and light-tailed distributions. Like MRPP (q.v. page 254), it is demonstrated in Sect. 6.16 of Chap. 6 that using raw measurements with $v = 1$ may be equally robust to outliers as using rank-order statistics with $v = 2$.

Table 5.1 Permutation generated subscripts for $N = 12$ with $n_1 = 3, n_2 = 4$, and $n_3 = 5$

Number	n_1			n_2				n_3				
1	1	2	3	4	5	6	7	8	9	10	11	12
2	1	2	3	4	5	6	8	7	9	10	11	12
3	1	2	3	4	5	7	8	6	9	10	11	12
4	1	2	3	4	6	7	8	5	9	10	11	12
5	1	2	3	5	6	7	8	4	9	10	11	12
6	1	2	3	4	5	6	9	7	8	10	11	12
7	1	2	3	4	5	7	9	6	8	10	11	12
8	1	2	3	4	6	7	9	5	8	10	11	12
9	1	2	3	5	6	7	9	4	8	10	11	12
10	1	2	3	4	5	8	9	6	7	10	11	12
⋮		⋮			⋮				⋮			
27,719	10	11	12	5	7	8	9	1	2	3	4	6
27,720	10	11	12	6	7	8	9	1	2	3	4	5

5.10 Subroutine PERMUT

In 1982 Berry published algorithm and non-recursive FORTRAN subroutine PERMUT to generate all permutations of multi-sets with fixed repetition numbers in Gray-code order [108]. Gray codes are a well-developed part of combinatorial algorithms, where they constitute systematic processes for complete enumeration of procedures such as the bootstrap and permutation tests [348]. Gray codes allow a listing of a set of combinatorial objects while making minimal changes at each step. Specifically, subroutine PERMUT enumerated the complete set of all permutations of N objects considered $n_1, n_2, \ldots, n_g$ at a time, i.e.,

$$\binom{N}{n_1\; n_2 \cdots n_g} = \frac{N!}{n_1!\; n_2! \cdots n_g!},$$

where N was partitioned into g unordered objects within each of the g groups, $N = n_1 + n_2 + \cdots + n_g$, and $n_1, n_2, \ldots, n_g$ were fixed for all permutations. If $g = 2$, then all combinations were enumerated, and also if $g = N$ and each $n_i = 1$, then all $N!$ values were enumerated [108].

Unlike other permutation routines, subroutine PERMUT enumerated the actual marks, rather than pointers to the marks (q.v. page 218), as was common at the time [247, 248, 783]. In addition, subroutine PERMUT converted the marks to integers, ready to be used as permuted subscripts to the observed values. This was an essential feature for implementation of exact permutation tests associated with g-sample procedures, such as MRPP (q.v. page 254). Table 5.1 illustrates a permutation structure generated by subroutine PERMUT for $N = 12$ observations with $g = 3$, $n_1 = 3$, $n_2 = 4$, and $n_3 = 5$, and where $12!/(3!\,4!\,5!) = 27{,}720$. In 1987 Berry published a version of subroutine PERMUT written in APL [109].

Fig. 5.6 Example data set for the F, randomization, Monte Carlo tests, and the moment-approximation procedure

S_1	S_2	S_3
43.75	46.00	50.50
50.50	61.75	68.50
43.75	46.00	64.00
	52.75	68.50
		50.50
		66.25

5.11 Moment Approximations and the *F* Test

In 1983 Berry and Mielke released an algorithm and FORTRAN program for computing the three exact finite population moments and the Pearson type III moment-approximation probability values for MRPP (qq.v. pages 254–265) [120]. A driver program read in the raw data and called the appropriate subroutines to (1) calculate the distance functions, (2) compute the exact values of μ_δ, σ_δ^2, and γ_δ, and (3) calculate the moment-approximation probability value based on the Pearson type III probability distribution [120].

Also in 1983, Berry and Mielke considered a moment-approximation permutation procedure as an alternative to the F test in a completely randomized analysis of variance design [121]. They noted that three alternatives to the traditional F test were often advocated when the requirements of equal variances or normally-distributed populations were questionable: (1) rely on the robustness of the F test when deviations from the assumptions were not severe, (2) substitute a Fisher–Pitman exact permutation test for the F test, or (3) employ a Monte Carlo resampling procedure. A fourth approach considered in this article was that of moment-approximation permutation procedures based on the Pearson type III probability distribution. Not considered in the article was any transformation of the original data, such as reduction of the raw data to ranks with its attendant loss of information.

An example analysis illustrated the moment-approximation approach. Consider $N = 13$ objects, randomly assigned to $g = 3$ experimental treatments (S_1, S_2, S_3) with $n_1 = 3$, $n_2 = 4$, and $n_3 = 6$, where $S_1 = \{43.75, 50.50, 43.75\}$, $S_2 = \{46.00, 61.75, 46.00, 52.75\}$, and $S_3 = \{50.50, 68.50, 64.00, 68.50, 50.50, 66.25\}$, as given in Fig. 5.6.

For these data, a conventional F test yielded an $F = 4.70$ and, with two and ten degrees of freedom, an approximate probability value of 0.0364. An exact F test based on

$$\binom{N!}{n_1!\, n_2!\, n_3!} = \frac{13!}{3!\; 4!\; 6!} = 60{,}060$$

permutations of the observed data yielded 2,470 of the 60,060 possible permutations possessing F-ratios greater than or equal to the realized F value of 4.70. Thus, the exact probability value computed from subroutine PERMUT was $2{,}470/60{,}060 = 0.0411$. A Monte Carlo test based on $L = 1{,}000$ resamplings yielded 31 F-ratios greater than or equal to the realized F value and an approximate probability value of $31/1{,}000 = 0.0310$. The moment-approximation procedure, based on the Pearson type III distribution, yielded $\delta = 113.231$, $\mu_\delta = 1{,}045.56$, and $\gamma_\delta = -1.3043$, with a moment-approximation probability value of 0.0377.

5.11.1 Additional Applications of MRPP

Along with the applications of MRPP discussed previously, a variety of other functions for MRPP have been developed. Specific examples of these data-dependent techniques include (1) the detection of autoregressive patterns; (2) analyses of asymmetric two-way contingency tables; (3) various measures of agreement for different types of data; (4) analyses of cyclic data, including circular and spherical data; and (5) both conventional and generalized runs test analyses [965, Chap. 3].

5.12 Mielke–Iyer and MRBP

In 1982 Paul Mielke and Hari Iyer published a class of procedures for randomized block analysis of variance designs that included the analysis of multivariate data [984].

H.K. Iyer

Hariharan (Hari) Kalahasty Iyer received his B.Sc. degree in mathematics from the University of Bombay (Mumbai) in 1970. Subsequently, he earned his M.S. and Ph.D. degrees in mathematics from Notre Dame University in 1972 and 1975, respectively. He was an instructor of mathematics at the University of Utah from 1975 to 1977, at which time he moved to Colorado State University to study experimental design with Professor Raj Chandra Bose. He received his second Ph.D., this time in statistics, from Colorado State University in 1980 and immediately accepted a position in the Department of Statistics at Colorado State University where he remained until his retirement in 2011.

In 1982 Mielke (q.v. page 249) and Iyer presented a class of permutation techniques for randomized block analysis of variance designs [984]. The class was specifically devised for analyses involving multivariate data. As Mielke and Iyer noted, many well-known techniques are special cases of this class, including (1) the permutation version of the classical univariate technique based on the F statistic for randomized blocks, (2) the Cochran Q test and the McNemar test, (3) the

Pearson product-moment correlation coefficient, (4) the matched-pairs t test, (5) the Friedman two-way analysis of variance for ranks test, and (6) the Spearman rank-order correlation and footrule measures.

Let $x'_{ij} = (x_{1ij}, \ldots, x_{rij})$ denote a vector of r commensurate response measurements associated with the ith treatment and jth block in a multivariate randomized block analysis of variance design. Then, the multi-response randomized block permutation procedures (MRBP) statistic is given by

$$\delta = \left[g \binom{b}{2} \right]^{-1} \sum_{i=1}^{g} \sum_{j<k} \Delta(x_{ij}, x_{ik}),$$

where $\sum_{j<k}$ is the sum over all j and k such that $1 \leq j < k \leq b$ and $\Delta(x, y)$ is the symmetric distance function value of the two r-dimensional vectors $x' = (x_1, \ldots, x_r)$ and $y' = (y_1, \ldots, y_r)$ in the r-dimensional Euclidean space. The choice of symmetric distance functions is given by

$$\Delta(x, y) = \left[\sum_{i=1}^{r} (x_i - y_i)^2 \right]^{v/2},$$

where $v > 0$. The null hypothesis (H_0) states that the distribution of δ assigns an equal probability to each of the

$$M = (g!)^b$$

possible allocations of the g r-dimensional response measurements to the g treatment positions within each of the b blocks. Consequently, the collection of r response measurements within each block yields g r-dimensional exchangeable random variables under H_0 (q.v. page 4). The δ statistic compares the within-group clustering of response measurements against the model specified by random allocation under H_0.

The exact MRBP probability values are analogous to the exact MRPP probability values described in Chap. 4. Like MRPP, the calculation of exact MRBP probability values becomes unreasonable when M exceeds, say, 10^9. Thus, resampling and Pearson type III moment-approximations are just as essential for MRBP as they are for MRPP. A resampling-approximation permutation test is based on L independent realizations of δ. Since each realization of δ requires a constant multiple of gb^2 operations, a resampling-approximation probability value requires a constant multiple of Lgb^2 operations. Alternatively, a Pearson type III moment-approximation depends on the exact mean, variance, and skewness of δ under H_0. If

$$\Delta(i, s; j, t) = \Delta\left(x_{is}, x_{jt}\right)$$

and

$$D(i,s;j,t) = \Delta(i,s;j,t) - g^{-1}\sum_{m=1}^{g}\Delta(m,s;j,t)$$
$$- g^{-1}\sum_{n=1}^{g}\Delta(i,s;n,t) + g^{-2}\sum_{m=1}^{g}\sum_{n=1}^{g}\Delta(m,s;n,t),$$

then μ_δ, σ_δ^2, and γ_δ are expressed as

$$\mu_\delta = \left[g^2\binom{b}{2}\right]^{-1}\sum_{s<t}\sum_{i=1}^{g}\sum_{j=1}^{g}\Delta(i,s;j,t),$$

$$\sigma_\delta^2 = \left[g\binom{b}{2}\right]^{-2}\frac{1}{g-1}\sum_{s<t}\sum_{i=1}^{g}\sum_{j=1}^{g}\left[D(i,s;j,t)\right]^2,$$

and

$$\gamma_\delta = \frac{\kappa_3}{\sigma_\delta^3},$$

using the definition of κ_3 given by

$$\kappa_3 = \left[g\binom{b}{2}\right]^{-3}\frac{1}{g-1}\left[H(g)+L(b)\right],$$

where $g \geq 2$, $b \geq 2$, and wherein, first, $H(g) = 0$ if $g = 2$, and

$$H(g) = \frac{g}{g-2}\sum_{s<t}\sum_{i=1}^{g}\sum_{j=1}^{g}\left[D(i,s;j,t)\right]^3$$

if $g \geq 3$, and, secondly, $L(b) = 0$ if $g = 2$, and

$$L(b) = \frac{6}{g-1}\sum_{s<t<u}\sum_{i=1}^{g}\sum_{j=1}^{g}\sum_{k=1}^{g}D(i,s;j,t)D(i,s;k,u)D(j,t;k,u)$$

if $b \geq 3$. Here, $\sum_{s<t<u}$ denotes the sum over all s, t, and u such that $1 \leq s < t < u \leq b$. Therefore, a Pearson type III probability value requires a constant multiple of g^3b^3 operations. The resampling and Pearson type III moment-approximations comparison described in Chap. 4 for MRPP also pertain to MRBP. Here, the execution time of a Pearson type III moment-approximation is roughly g^2b/L that of a resampling-approximation. In 1983, Iyer, Mielke, and Berry published a computer program for computing finite population parameters and approximate probability values for MRBP using the Pearson type III distribution [676]. In 1991 Tracey and Khan provided the fourth exact moment of the test statistic in an effort to obtain a better approximating function [1367] and, in a separate article simplified the 35 symmetric functions required for the fourth moment. In this second article, Tracey and Khan also evaluated the power of MRBP tests with $v = 1$ and $v = 2$ using both three and four moments for the Laplace, Cauchy, and normal distributions [1368].

5.13 Relationships of MRBP to Other Tests

MRBP are related to a number of conventional tests and measures. If $v = 2$ and $r = 1$, then the functional relationships between the randomized-block F-ratio test statistic and δ are given by

$$F = \frac{(b-1)[2SS_T - g(b-1)\delta]}{g(b-1)\delta - 2SS_B} \qquad \text{and} \qquad \delta = \frac{2SS_B + (b-1)SS_T}{(b-1)g(F+b-1)},$$

where SS_T, the corrected total sum of squares, is given by

$$SS_T = \sum_{i=1}^{g} \sum_{j=1}^{b} x_{ij}^2 - SS_M,$$

the block sum of squares, SS_B, is given by

$$SS_B = \left\{ \sum_{j=1}^{b} \left[\frac{\left(\sum_{i=1}^{g} x_{ij}\right)^2}{g} \right] \right\} - SS_M,$$

and SS_M, the correction factor, is given by

$$SS_M = \frac{\left(\displaystyle\sum_{i=1}^{g} \sum_{j=1}^{b} x_{ij} \right)^2}{bg}.$$

Thus, F and δ are equivalent under the null hypothesis since both SS_T and SS_B are invariant relative to the $(g!)^b$ permutations of the response measurements.

If $r = 1$, $b = 2$, and each x_{ij} is either 0 or 1, δ is equivalent to McNemar's test for change [916], and if $r = 1$, $b > 2$, and each x_{ij} is either 0 or 1, then δ is equivalent to Cochran's Q test statistic [259].

If $g = 2$, $r = 1$, $x_{1j} = -x_{2j} = x_j$ and $|x_j| > 0$ for $j = 1, \ldots, b$, then the test based on δ is equivalent to an extended class of permutation techniques for matched-pairs data [945].

If $v = 2$, $r = 1$, and the response measurements for each block are replaced by their corresponding ranks, then the test based on δ is equivalent to the Friedman two-way analysis of variance for ranks, the Kendall coefficient of concordance, and the Wallis correlation ratio for ranked data [485, 486, 734, 739, 1411].

If $v = 2$, $r = 1$, $b = 2$, and the response measurements for each block are replaced by their corresponding ranks, then $1 - \delta/\mu_\delta$ is Spearman's rank-order correlation coefficient [1300], where $\mu_\delta = (g^2 - 1)/6$. And, if $v = r = 1$, $b = 2$, and the response measurements for each block are again replaced by their corresponding ranks, then the test based on δ given by $1 - \delta/\mu_\delta$ is the Spearman footrule statistic where $\mu_\delta = (g^2 - 1)/(3g)$ [1301].[23]

If $v = 2$, $b = 2$, and $r = 1$, then the functional relationship between δ and the Pearson product-moment correlation coefficient, R, is given by

$$R = \frac{\mu_\delta - \delta}{2S_1S_2} \qquad \text{and} \qquad \delta = \mu_\delta - 2RS_1S_2,$$

where

$$R = \frac{\displaystyle\sum_{i=1}^{g} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{gS_1S_2},$$

$$\mu_\delta = S_1^2 + S_2^2 + (\bar{x}_1 - \bar{x}_2)^2,$$

$$\bar{x}_j = \tfrac{1}{g} \sum_{i=1}^{g} x_{ij},$$

and

$$S_j^2 = \frac{\displaystyle\sum_{i=1}^{g} \left(x_{ij} - \bar{x}_j\right)^2}{g}$$

for $j = 1$ and 2. R and δ are equivalent under the null hypothesis since $\bar{x}_1$, $\bar{x}_2$, S_1, and S_2 are invariant relative to the $(g!)^2$ permutations of the response measurements.

[23] As noted by John W. Whitfield in 1950 [1444], Spearman argued that the process of squaring the deviations, i.e., $v = 2$, increased the error of the correlation measure and repeated this argument in various publications; see especially three articles by Pitman in 1904, 1906, and 1910 [1300–1302].

5.14 Kappa and the Measurement of Agreement

A number of statistical problems require the measurement of agreement, rather than association or correlation, between two or more independent raters [15].[24] One of the most popular indices of agreement is Cohen's kappa (κ) introduced in 1960 as an index of inter-rater agreement for categorical (nominal level) variables [263].

J. Cohen

Jacob Cohen was a New Yorker. He was born in New York City, educated in New York City, worked in New York City, and died in New York City. Cohen entered City College of New York (CCNY) at the age of 15 and put in a stint with Army Intelligence in France before graduating from CCNY in 1947. He went on to earn an M.A. and Ph.D. in clinical psychology from New York University (NYU) in 1948 and 1950, respectively. In 1949 he was appointed as an Instructor at New York University and promoted to Professor 10 years later.

As Kevin Murphy wrote in 1998, Cohen made three major contributions to quantitative methods. First, Cohen's kappa was cited in his Distinguished Lifetime Contribution Award as "the gold standard for the measurement of agreement between categorical judgments"; second, Cohen championed the use of multiple regression as a general data-analytic framework; and third, Cohen's work on statistical power analysis changed the way researchers think about tests of significance and measures of effect size [1015]. Jacob Cohen retired from New York University in 1993 after 44 years of service and passed away on 20 January 1998 at the age of 74 after a lengthy illness.

As developed by Cohen, kappa was a chance-corrected measure of agreement between two independent raters, each rating n observations on a nominal level scale of measurement with c categories. More specifically, kappa was proposed as a chance-corrected measure of agreement to discount the proportion of agreement by the expected level of agreement, given the observed marginal frequency distributions of the raters' responses and the assumption that the rater reports are statistically independent [60, p. 4]. Kappa is equal to 1 when perfect agreement between two independent raters occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than that expected by chance. Writing in 2008, von Eye and von Eye would be more precise, describing kappa as a measure of the degree to which the agreement cells on the principal diagonal contain more cases than expected under the model of rater independence and noting that kappa only attains a value of 1 when the marginal frequency totals

[24] "Raters" are variously termed "judges" or "observers" in the agreement literature.

Table 5.2 Example 5 × 5 cross-classification table with cell proportions

	Column					
Row	1	2	3	4	5	Row total
1	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	$p_{1.}$
2	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	$p_{2.}$
3	p_{31}	p_{32}	p_{33}	p_{34}	p_{35}	$p_{3.}$
4	p_{41}	p_{42}	p_{43}	p_{44}	p_{45}	$p_{4.}$
5	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	$p_{5.}$
Column total	$p_{.1}$	$p_{.2}$	$p_{.3}$	$p_{.4}$	$p_{.5}$	$p_{..}$

are the same [1401, p. 313]. In 1988 Berry and Mielke generalized Cohen's kappa to ordinal and interval levels of measurement, as well as to multiple raters [133].

Consider two independent raters who classify each of n observations into one of c a priori nominal categories. The resulting classifications can be displayed in a $c \times c$ cross-classification (agreement) table, such as that in Table 5.2, with proportions for cell entries. In the notation of Table 5.2, Cohen's kappa is given by

$$\kappa = \frac{P_\mathrm{o} - P_\mathrm{e}}{1 - P_\mathrm{e}}, \tag{5.1}$$

where

$$P_\mathrm{o} = \sum_{i=1}^{c} p_{ii} \quad \text{and} \quad P_\mathrm{e} = \sum_{i=1}^{c} p_{i.}p_{.i} .$$

In this formulation, P_o is the observed proportion of observations on which the raters agree, P_e is the proportion of observations for which agreement is expected by chance, $P_\mathrm{o} - P_\mathrm{e}$ is the proportion of agreement beyond what is expected by chance, $1 - P_\mathrm{e}$ is the maximum possible proportion of agreement beyond what is expected by chance, and the kappa coefficient, κ, is the proportion of agreement between the two independent raters after chance agreement has been removed.

Alternatively, let $\delta = 1 - P_\mathrm{o}$ represent the observed proportion of disagreement and $\mu_\delta = 1 - P_\mathrm{e}$ represent the expected proportion of disagreement. Then, substitution into Eq. (5.1) and simplification yields

$$\kappa = 1 - \frac{\delta}{\mu_\delta}. \tag{5.2}$$

Thus, Cohen's kappa may be interpreted as a ratio of measures of distance, or disagreement, between the two raters, where the distance between the raters is measured by counting up a series of zeroes and ones [825]. In this form, kappa is a measure of agreement based on the proximity of the classifications, which is measured by the Euclidean distance between the classifications of the two independent raters [133].

Table 5.3 Alternative representation of Table 5.2

	Observation						
Block	1	2	3	4	5	$\cdots$	n
1	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$	$\mathbf{x}_{14}$	$\mathbf{x}_{15}$	$\cdots$	$\mathbf{x}_{1n}$
2	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$	$\mathbf{x}_{24}$	$\mathbf{x}_{25}$	$\cdots$	$\mathbf{x}_{2n}$

An alternative representation of Table 5.2 was given by Berry and Mielke, which lends itself to analysis by Eq. (5.2) and is presented in Table 5.3. The depiction in Table 5.3 is constructed in the context of a multivariate randomized block design with n observations, two blocks corresponding to the two raters, and the polytomous variable of Table 5.2 represented by a $c \times 1$ vector $\mathbf{x}$ where the ith element, corresponding to the ith of the c categories, is set to $2^{-1/2}$, and where the remaining $c - 1$ elements of $\mathbf{x}$ are set to zero. The choice of the constant $2^{-1/2}$ is simply to ensure that the distance between any two vectors will be zero if the classifications agree, and one if the classifications disagree [133].

In this second formulation, δ is given by

$$\delta = \frac{1}{n} \sum_{i=1}^{n} \Delta \left(\mathbf{x}_{1i} , \, \mathbf{x}_{2i} \right) , \tag{5.3}$$

where

$$\Delta \left(\mathbf{x}_{1i} , \, \mathbf{x}_{2i} \right) = \left[\sum_{k=1}^{c} \left(x_{1ik} - x_{2ik} \right)^2 \right]^{1/2} ,$$

and x_{rik} denotes the kth element of vector $\mathbf{x}_{ri}$ with $r = 1, 2$ for blocks 1 and 2, and $i = 1, \ldots, n$ for observations 1 through n. Then μ_δ is the expected proportion of disagreement, which is defined as

$$\mu_\delta = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta \left(\mathbf{x}_{1i} , \, \mathbf{x}_{2i} \right) , \tag{5.4}$$

where

$$\Delta \left(\mathbf{x}_{1i} , \, \mathbf{x}_{2j} \right) = \left[\sum_{k=1}^{c} \left(x_{1ik} - x_{2jk} \right)^2 \right]^{1/2} .$$

The two representations of κ in Eqs. (5.1) and (5.2) are equivalent mathematical formulations of the same structure. The advantage of the second representation of κ given in Eq. (5.2) is that it simplifies extension and generalization to multiple raters and to higher levels of measurement [133].

Given the multivariate randomized block representation in Table 5.3 and Eqs. (5.3) and (5.4), Berry and Mielke proposed two extensions of Cohen's kappa. First, they showed that kappa could easily be extended to levels of measurement other than the nominal level and, second, kappa could readily be expanded to more than two raters.

5.14.1 Extensions to Interval and Ordinal Data

The construction of $\Delta\left(\mathbf{x}_{1i}, \mathbf{x}_{2i}\right)$ makes an extension of kappa to interval measurements straightforward. For interval data, $\mathbf{x}$ is simply a vector of 1 to c measurements. Instead of a rater assigning each observation to one of c categories, the rater assigns a score, or a vector of c scores in the multivariate case, to each observation. In this case, kappa measures the degree to which the two raters agree on their scoring, above and beyond what is expected by chance. For ordinal data, $\mathbf{x}$ is a vector of 1 to c ranks, where a rater assigns a rank, or a vector of c ranks in the multivariate case, to each observation.

Although the formulae for δ and μ_δ given in Eqs. (5.3) and (5.4) are unaffected by ordinal or interval measurements, the extension of Cohen's kappa to higher levels of measurement is sufficiently general to require a distinguishing symbol. Berry and Mielke proposed

$$\Re = 1 - \frac{\delta}{\mu_\delta} \tag{5.5}$$

for any level of measurement, where $\Re$ is equivalent to Cohen's kappa coefficient as defined in Eq. (5.1) with measurements made at the nominal level of measurement [133].

5.14.2 Extension of Kappa to Multiple Raters

A simple modification to the computation of δ and μ_δ given in Eqs. (5.3) and (5.4) generalizes $\Re$, as given in Eq. (5.5), to measure agreement among multiple raters. Thus, δ as given in Eq. (5.3) was redefined by Berry and Mielke [133] for multiple raters as

$$\delta = \left[n \binom{b}{2} \right]^{-1} \sum_{i=1}^{n} \sum_{r<s} \Delta\left(\mathbf{x}_{ri}, \mathbf{x}_{si}\right), \tag{5.6}$$

where

$$\Delta\left(\mathbf{x}_{ri}, \mathbf{x}_{si}\right) = \left[\sum_{k=1}^{c} \left(x_{rik} - x_{sik}\right)^2 \right]^{1/2},$$

b is the number of blocks (i.e., raters), and $\sum_{r<s}$ is the sum over all r and s such that $1 \leq r < s \leq b$. The reformulation of μ_δ was given by

$$\mu_\delta = \left[n^2 \binom{b}{2} \right]^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r<s} \Delta\left(\mathbf{x}_{ri}, \mathbf{x}_{sj}\right), \tag{5.7}$$

where

$$\Delta\left(\mathbf{x}_{ri}, \mathbf{x}_{sj}\right) = \left[\sum_{k=1}^{c} \left(x_{rik} - x_{sjk}\right)^2 \right]^{1/2}.$$

If $b = 2$, Eqs. (5.6) and (5.7) reduce to Eqs. (5.3) and (5.4), respectively.

Since $\Re$ is simply a linear transformation of δ, the test of significance for $\Re$ is the test of significance for δ. As the b blocks are specified, the randomization associated with a randomized block design is confined to all permutations of the n observations within each block. Under the null hypothesis, each of the $M = (n!)^b$ permutations has an equal probability of occurrence; viz., $1/M$. Berry and Mielke utilized a moment-approximation approach based on the Pearson type III distribution as follows: if δ_j denotes the jth value among the M possible values of δ, then the exact mean, variance, and skewness of δ, under the null hypothesis, are given by

$$\mu_\delta = \frac{1}{M} \sum_{j=1}^{M} \delta_j,$$

$$\sigma_\delta^2 = \frac{1}{M} \sum_{j=1}^{M} \delta_j^2 - \mu_\delta^2,$$

and

$$\gamma_\delta = \left(\frac{1}{M} \sum_{j=1}^{M} \delta_j^3 - 3\mu_\delta \sigma_\delta^2 - \mu_\delta^3 \right) \Big/ \sigma_\delta^3 ,$$

respectively. If δ_{o} denotes the observed value of δ, then the approximate probability value based on the Pearson type III distribution is given by

$$P(\delta \leq \delta_{\text{o}} \mid H_0) \doteq \int_{-\infty}^{T_{\text{o}}} f(y)\, dy,$$

where

$$T_{\text{o}} = \frac{\delta_{\text{o}} - \mu_\delta}{\sigma_\delta}.$$

5.14.3 Limitations of Kappa

Kappa is well known as a marginal-dependent measure of agreement and is often criticized on this basis; see for example, articles by Brennan and Prediger in 1981 [210], Maclure and Willett in 1987 [864], Thompson and Walter in 1988 [1359], Zwick in 1988 [1498], Guggenmoos-Holzmann in 1993 and 1995 [560, 561], May in 1994 [909], and Agresti in 2002 [4]. The problem is that there are two sources of disagreement, differences in thresholds and differences in construction of the underlying continuous scale, and it is inherently impossible to represent them by a single number, as noted by Brennan and Hays in 1992 [209] and Hutchinson in 1993 [671]. It shares this characteristic with Pearson's chi-squared statistic and the product-moment correlation coefficient for cross classifications. Thus, kappa cannot approximate its maximum value of 1.00 when the marginal frequency distributions in an agreement classification table are not uniform, as noted by von Eye and von Eye in 2008 [1401]. However, kappa will attain its maximum value of 1.00 when the probability for all disagreement cells is zero; consequently, kappa shows no marginal dependency under conditions of perfect agreement [1401]. As Brennan and Prediger noted in 1981: "[i]t is evident that indiscriminate use of coefficient kappa without modification *may* lead to dramatically incorrect conclusions about the proportion of maximum possible agreement evident in a set of data" [210, p. 698].[25]

Cohen's kappa is considered the gold standard among agreement coefficients and is interpreted as the proportionate increase in rater agreement above and beyond what can be expected by chance alone, where chance is defined as the level of agreement expected if the raters had a known base rate for the objects under study and randomly assigned cases corresponding to the base rate; see also a 1997 article on the assessment of reliability by Meyer [929]. This definition of chance has been referred to by Brennan and Prediger [210] and Umesh, Peterson, and Sauber [1384] as the "fixed marginals" model because the marginal distributions of category assignment are assumed to be known a priori. The problem with the fixed-marginals approach is that it does not give the raters credit for assignments that are independently agreed upon and reflected in the marginal distributions. Thus, as noted by Brennan and Prediger [210], Hanley [584], and Zwick [1498], Cohen's kappa statistic penalizes the raters by using the base rate to define the chance agreement level the raters must surpass. For example, consider two raters and two categories, A and B. If the two raters both feel that the base rate in the population for category A is 0.10 and each judge randomly assigns 10 % of the cases to category A, then by chance alone the percentage agreement between the two raters is $(0.10)(0.10) + (0.90)(0.90) = 0.82$, and the observed agreement between the two raters must exceed 0.82 for the computed value of kappa to be greater than zero.

Cohen's kappa is extremely sensitive to the base-rate phenomenon. Because the maximum value that kappa can attain is constrained by differences between the marginal distributions of the two raters, as the base rate moves away from

[25] Emphasis in the original.

Fig. 5.7 Classification of $n = 10$ objects into $c = 4$ categories by $b = 2$ raters

	Rater	
Object	I	II
1	A	A
2	A	A
3	A	A
4	A	A
5	A	A
6	A	A
7	A	A
8	A	A
9	A	A
10	B	B

Fig. 5.8 An agreement classification table for the ratings of the $n = 10$ objects listed in Fig. 5.7

	A	B	C	D	
A	9	0	0	0	9
B	0	1	0	0	1
C	0	0	0	0	0
D	0	0	0	0	0
	9	1	0	0	10

the point of maximum variability a small disagreement between the raters can cause the kappa value to decline dramatically, as noted by Meyer in 1997 [929]. On the other hand, some researchers have argued that this is appropriate as kappa is a true reliability statistic; see for example, articles by Cohen in 1960 [263], Shrout, Spitzer, and Fleiss in 1987 [1270], and Bartko in 1991 [77]. That is to say, as true score variability in the group becomes more restricted, a fixed amount of disagreement plays an increasingly larger role in observed score variability, so calculated reliability coefficients decline in value. In 1960 Cohen noted that it is perfectly reasonable and, in fact, desirable to use a summary agreement measure that is sensitive to both aspects of agreement: item-by-item agreement as reflected in the main diagonal of the agreement matrix, and symmetry between the marginal distributions [263]. Also, Spitznagel and Helzer protested against even providing base-rate information, arguing that it defeats the purpose of a single measure of reliability [1311].

That the magnitude of kappa is conditional on the marginal frequency distributions can easily be demonstrated. Consider $n = 10$ objects classified into $c = 4$ a priori, mutually-exclusive, nominal categories $\{A, B, C, D\}$ by $b = 2$ independent raters. Also, for simplicity, assume that both raters classified all ten objects into just two of the four categories, A and B, as illustrated in Fig. 5.7. The classification data in Fig. 5.7 have been cross-classified into an agreement classification table in Fig. 5.8. It is readily apparent from even a casual visual inspection of Figs. 5.7 and 5.8 that the raters are in perfect agreement and kappa should therefore be equal to 1.00, given that the marginal frequency distributions are uniform, i.e., $\{9, 1, 0, 0\}$

Fig. 5.9 Second classification of $n = 10$ objects into $c = 4$ categories by $b = 2$ raters

	Rater	
Object	I	II
1	A	A
2	A	A
3	A	A
4	A	A
5	A	A
6	A	A
7	A	A
8	A	A
9	A	A
10	A	B

and $\{9, 1, 0, 0\}$. Thus, for the data in Fig. 5.7 the observed proportion of agreement on the principal diagonal is

$$P_{\text{o}} = \sum_{i=1}^{c} p_{ii} = \left(\frac{9}{10}\right) + \left(\frac{1}{10}\right) + \left(\frac{0}{10}\right) + \left(\frac{0}{10}\right) = \frac{10}{10} = 1.00,$$

the expected proportion of agreement on the principal diagonal is

$$\begin{aligned} P_{\text{e}} = \sum_{i=1}^{c} p_{i.}p_{.i} &= \left(\frac{9}{10}\right)\left(\frac{9}{10}\right) + \left(\frac{1}{10}\right)\left(\frac{1}{10}\right) + \left(\frac{0}{10}\right)\left(\frac{0}{10}\right) + \left(\frac{0}{10}\right)\left(\frac{0}{10}\right) \\ &= 0.81 + 0.01 + 0.00 + 0.00 = 0.82, \end{aligned}$$

and following Eq. (5.1),

$$\kappa = \frac{1.00 - 0.82}{1 - 0.82} = \frac{0.18}{0.18} = 1.00.$$

Also note that the percentage agreement is $P_{\text{o}}(100) = 1.00(100) = 100\,\%$, i.e., both raters agreed on the classification of all ten objects.

Next, consider the same $n = 10$ objects classified into the same $c = 4$ a priori, mutually-exclusive, nominal categories by the same $b = 2$ independent raters. As before, assume that both raters classified all ten objects into just two of the four categories, A and B, as illustrated in Fig. 5.9.

The classification data listed in Fig. 5.9 have been cross-classified into an agreement classification table in Fig. 5.10. Here the two raters are in perfect agreement except for the classification of Object 10, where Rater I classified Object 10 into Category A and Rater II classified Object 10 into Category B. Note that in this case the marginal frequency distributions in Fig. 5.10 are no longer uniform, i.e., $\{10, 0, 0, 0\}$ and $\{9, 1, 0, 0\}$, and the magnitude of kappa will therefore be

Fig. 5.10 An agreement classification table for the ratings of the $n = 10$ objects listed in Fig. 5.9

	A	B	C	D	
A	9	1	0	0	10
B	0	0	0	0	0
C	0	0	0	0	0
D	0	0	0	0	0
	9	1	0	0	10

restricted by the nonuniform marginal frequency distributions. For the classified data in Fig. 5.10, the observed proportion of agreement on the principal diagonal is

$$P_{\text{o}} = \frac{9}{10} + \frac{0}{10} + \frac{0}{10} + \frac{0}{10} = \frac{9}{10} = 0.90,$$

the expected proportion of agreement on the principal diagonal is

$$\begin{aligned} P_{\text{e}} &= \left(\frac{10}{10}\right)\left(\frac{9}{10}\right) + \left(\frac{0}{10}\right)\left(\frac{1}{10}\right) + \left(\frac{0}{10}\right)\left(\frac{0}{10}\right) + \left(\frac{0}{10}\right)\left(\frac{0}{10}\right) \\ &= 0.90 + 0.00 + 0.00 + 0.00 = 0.90, \end{aligned}$$

and following Eq. (5.1),

$$\kappa = \frac{0.90 - 0.90}{1 - 0.90} = 0.00.$$

However, the percentage agreement is $P_{\text{o}} = 0.90(100) = 90\,\%$, i.e., the two raters agreed on $90\,\%$ (nine of ten) of the classifications of the ten objects.

In these carefully constructed examples, a single small shift in the classification of one object from a B to an A by one rater resulted in a dramatic change, dropping the magnitude of kappa from $\kappa = 1.00$ down to $\kappa = 0.00$, all due to the restriction on kappa imposed by the nonuniform marginals. Although in this extreme example the simple percentage of agreement appears preferable, many researchers acknowledge that chance-corrected measures, such as kappa, are an improvement over non-chance-corrected measures of agreement, as the latter do not adjust for the fact that a certain amount of agreement could have occurred due to chance alone; see for example, articles by Cicchetti and Feinstein in 1990 [256], Feinstein and Cicchetti in 1990 [423], Byrt in 1992 [231], Byrt, Bishop, and Carlin in 1993 [232], and Graham in 1995 [543].

Finally, it should be noted that while the two examples are based on unweighted kappa with two raters, the same marginal restraints on the magnitude of kappa hold for multiple raters ($b > 2$) and for weighted kappa with either linear or quadratic weighting. In general, the problem with Cohen's kappa is referred to as the "base-rate" problem. For further discussions of the base-rate problem, see articles by Carey and Gottesman in 1978 [242]; Brennan and Prediger in 1981 [210], Soeken and Prescott in 1986 [1291]; Thompson and Walter in 1988 [1359]; Zwick in 1988 [1498], Ker in 1991 [750]; Seigel, Padgor, and Remaley in 1992 [1244]; Hutchinson in 1993 [671]; Agresti and Ghosh in 1995 [5]; Meyer in 1997 [929]; Banerjee, Capozzoli, McSweeney, and Sinha in 1999 [60]; Nelson and Pepe in 2000 [1029]; Hsu and Field in 2003 [660]; Kundel and Polansky in 2003 [781]; Martín Andrés and Marzo in 2004 and 2005 [901, 902]; and de Mast in 2007 [333].

5.14.4 Relationships Between $\Re$ and Existing Measures

Relationships between $\Re$, as given in Eq. (5.5), and existing measures were detailed by Berry and Mielke in 1988 [133]. They explained that the generalization of Cohen's kappa to multiple raters for nominal-level data is the special case of $\Re$ when the distance space is restricted to a c-dimensional simplex, i.e., a distance space consisting of c distinct points where the distance between any two points is unity and the distance between any two coincident points at any one of the c positions is zero [133]. In this context, Berry and Mielke noted that Cohen's kappa [263] is the special case of $\Re$ when $b = 2$, the measure of agreement corresponding to Cochran's Q statistic [259] is the special case of $\Re$ when $c = 2$, and the measure of agreement corresponding to McNemar's test for change [916] is the special case of $\Re$ when $b = c = 2$ [941].

When the distance space is a one-dimensional Euclidean space and the observations are rank-order statistics of the n observations associated with each of the b raters, then the measure of agreement corresponding to Spearman's footrule, R, is a special case of $\Re$ when $b = 2$ [938].

C.E. Spearman

The life of Charles Edward Spearman spanned an interesting era of technological innovation. He was born on 10 September 1863, 13 years before the telephone was invented in March of 1876 by Alexander Graham Bell and 16 years before the carbon electric light bulb was invented in October of 1879 by Thomas Alva Edison. Charles Edward Spearman F.R.S. came to an untimely death on 17 September 1945 at age 82 after falling from a window of his hospital room in London just two months after the detonation of the atomic device at the Trinity site near Alamogordo, New Mexico, ushered in the Atomic Age on 16 July 1945. A historical note for classical music buffs: Pietro Mascagni, best known for his one-act opera *Cavalleria Rusticana* (Rustic Chivalry) was also born in 1863 (7 December) and died in 1945 (2 August).

(continued)

At age 19, Spearman graduated from Leamington College and his family secured a commission for him in the British Army with the Royal Munster Fusiliers. He served for 11 years with the 2nd Battalion in India and 2 years in Burma (now, Myanmar). In December 1896, while in the British Army he completed a 2-year course at the Army Staff College in Camberley, Surrey, from which he gained the coveted qualification of "Passed Staff College" [847]. At age 32 Spearman resigned his commission as a Major in the British Army and entered the University of Leipzig to study experimental psychology with Wilhelm Wundt. Four years later, Spearman was recalled to the British Army as Staff Officer for Guernsey during the Boer War in Africa. At the age of 40 he retired from the British Army for a second time and returned to Leipzig where he earned his Ph.D. in experimental psychology in 1906 at the age of 43 under the direction of Wundt [246, 684, 1119, 1303].[26]

In his autobiography, Spearman reflected upon his years in the British military:

> I committed the mistake of my life. Having no vocational advisor to assist me, I gave myself up to the youthful delusion that life is long. The problems which were now baffling me might perhaps, I thought, succumb to ripened experience. Following the illustrious example of René Descartes—not to mention Socrates and Plato—I decided to turn to a short spell of military service. This diversion of activity was, for one reason and another, allowed to spin out far longer than originally anticipated; it lasted until 1897. And for these almost wasted years I have since mourned as bitterly as ever Tiberius did for his lost legions [1303, p. 300].

After further study in Germany under Oswald Külpe at the University of Würzburg and George Elias Müller at the University of Göttingen, where he also attended the lectures of Edmund Husserl, Spearman returned to England in 1907 to a post at University College, London, where he remained until his retirement in 1931, having been appointed Grote Professor of Mind and Logic, and served as head of the Department of Psychology and President of the British Psychological Society [1303]. Spearman was elected Fellow of the Royal Society in 1924. Other than the 1904 and 1906 articles on rank-order correlation, Spearman is best remembered today as the father of testing theory, for the identification of a general factor of individual differences in mental abilities, and for his early work on factor analysis [230, 845, 846, 1459, 1497, pp. 243–245].

Spearman [1300, 1301] offered several coefficients for measuring the correlation between sets of ranked data, and it has been stated that it is not readily apparent

[26] Wilhelm Maximilian Wundt is known as "the founding father of experimental psychology" and many psychologists trace their academic legacy to Wundt, perhaps because he produced 186 Ph.D. students during his long career at the University of Leipzig. [755, 1497, pp. 141–143]

exactly which coefficient he meant to be associated with the term “footrule” [352, 844, 1109]. However,

$$R = 1 - \frac{3\sum_{i=1}^{n} \left| x_i - y_i \right|}{n^2 - 1}, \tag{5.8}$$

where x_i and y_i for $i = 1, \ldots, n$ denote n paired ranks, is widely accepted as the formula for Spearman’s footrule measure [347, 477, 736, 1217, 1327, 1387]. In addition, Spearman is explicit in his 1910 publication on “Correlation calculated from faulty data” that Eq. (5.8) is the formula for the footrule.[27] Spearman introduced the footrule as an easy but precise method of measuring the correlation between two rankings in 1906 [1300, 1301]. More specifically, referring to Pearson’s product-moment correlation coefficient as r, Spearman wrote that he had

> [e]xpressly entitled [sic] the measure as ‘footrule’ as lying half-way between the r method with its complications (which I likened to an ‘elaborate micrometer’ and judgment without mathematical method as all (which I compared to a ‘mere glance of the eye’). … R’s chief mission is merely to gain quickly an approximate valuation of r [1302, p. 286].

Unlike other measures of rank-order correlation, the footrule does not norm properly between the limits of -1 and $+1$. The footrule attains a maximum value of $+1$ when each x_i is identical to y_i for $i = 1, \ldots, n$ and no ties are present. However, if $y_i = n - x_i + 1$, then $R = -0.5$ when n is odd and

$$R = -0.5\left(1 + \frac{3}{n^2 - 1}\right)$$

when n is even [736]. Consequently, R does not attain a minimum value of -1 except for the trivial case when $n = 2$. Karl Pearson criticized the footrule on this basis in 1907 [1109] and Maurice Kendall explicitly pointed to this apparent lack of proper norming as a defect in the footrule as late as 1962 [736, p. 33]. Spearman, recognizing that negative values of R did not represent inverse correlation, actually suggested that “it is better to treat every correlation as positive” [1300, pp. 87–88], and writing later in 1910 he dismissed the problem entirely, stating that in all cases the negative correlation is less than its own probable error [1302, p. 285].

It can easily be shown that the footrule is a chance-corrected measure of agreement and not a measure of correlation[28] since it takes the classic form of a chance-corrected measure of agreement,

[27] Spearman’s test statistic, R, is based on the absolute differences between x_i and y_i, $i = 1, \ldots, n$. Thus, the absolute distance between two rank vectors is often referred to as the “footrule distance.” When the variables are quantitative, the absolute distance is known as a “city-block metric” or “Manhattan distance.”

[28] It is not generally recognized that under special conditions Spearman’s rank-order correlation coefficient is also a chance-corrected measure of agreement. When x and y consist of ranks from 1 to n with no ties, or x includes tied ranks and y is a permutation of x, then Spearman’s rank-

$$\text{agreement} = 1 - \frac{\text{observed agreement}}{\text{expected agreement}}$$

[772, p. 140] where, with no tied ranks, the expected agreement is given by $(n^2 - 1)/3$. As a chance-corrected measure of agreement, R is zero under chance conditions, unity when agreement is perfect, and negative under conditions of disagreement. The fact that the footrule does not norm to -1 with complete inversion of the rankings is therefore recast as a previously undocumented attribute of the footrule rather than a defect, placing the footrule firmly into the family of chance-corrected measures of agreement; see a 1997 paper on this topic by Berry and Mielke in *Psychological Reports* [147].

As originally formulated, Spearman's footrule was limited to fully ranked data and did not accommodate tied ranks. As Berry and Mielke [147] explained in 1997, let

$$\delta = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - y_i \right| ,$$

$$\mu_\delta = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| x_i - y_j \right| ,$$

and let

$$\Re = 1 - \frac{\delta}{\mu_\delta}$$

denote the general measure of the relationship between two sets of ranks that is not limited to untied ranks. If no ties exist in either x_i or y_j for $i, j = 1, \ldots, n$, then Spearman's footrule is given by

$$R = 1 - \frac{3n\delta}{n^2 - 1} .$$

In this way, the footrule is generalized to include tied ranks on x and y, and R is shown to be a special case of $\Re$ when no ties are present [147]. $\Re$, like R, is a chance-corrected measure of agreement since $\mathrm{E}[\delta] = \mu_\delta$.

When both x and y consist entirely of untied ranks from 1 to n and y is a permutation of x, then it is possible to determine the probability of an observed R under the null hypothesis that all of the $n!$ orderings of either the x or y values are equally likely. If

order correlation coefficient is both a measure of correlation and a chance-corrected measure of agreement [772, p. 144].

$$D = \sum_{i=1}^{n} \left| x_i - y_i \right| = n\delta ,$$

then since R is a linear transformation of D, the probability of an observed value of D is the probability of an observed value of R [147, p. 842]. Tables of the exact cumulative distribution function (cdf) of D were published by Ury and Kleinecke for $2 \leq n \leq 10$; in addition, approximate results based on Monte Carlo methods were provided for $11 \leq n \leq 15$ [1387]. Franklin [477] reported the exact cdf of D for $11 \leq n \leq 18$, and both Franklin [477] and Ury and Kleinecke [1387] discussed the rate of convergence to an approximating normal distribution and the use of a continuity correction to be applied to the cdf of D. Salama and Quade [1217] used Markov chain properties to obtain the exact cdf of D for $4 \leq n \leq 40$, corrected some tabled values in Franklin [477], and further investigated the adequacy of approximations to the distribution of D.

In 1998 Berry and Mielke generalized Spearman's footrule to encompass $b \geq 2$ sets of ranks [148]. Let

$$\delta = \left[n \binom{b}{2} \right]^{-1} \sum_{r<s} \sum_{i=1}^{n} \left| x_{ri} - x_{si} \right|$$

denote an average distance function based on all $\binom{b}{2}$ possible paired absolute differences among values of the b rankings, and let

$$\mu_\delta = \left[n^2 \binom{b}{2} \right]^{-1} \sum_{r<s} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| x_{ri} - x_{sj} \right|$$

denote the expected value of δ, where b is the number of rankings, n is the number of objects, and $\sum_{r<s}$ is the sum over all r and s such that $1 \leq r < s \leq b$. Then

$$\Re = 1 - \frac{\delta}{\mu_\delta} \tag{5.9}$$

is a chance-corrected measure of the agreement among the b rankings which is not limited to untied ranks. Note that in the case where $b = 2$ and there are no tied ranks, Eq. (5.9) reduces to Eq. (5.8), i.e., Spearman's footrule.

Also, if the distance space is comprised of squared Euclidean distances where

$$\Delta \left(\mathbf{x}_{ri}, \mathbf{x}_{sj} \right) = \sum_{k=1}^{c} \left(x_{rik} - x_{sjk} \right)^2 ,$$

then if the number of nominal categories, c, is equal to one and the n observations are rank-order statistics associated with each of the b raters, Spearman's rank-order correlation coefficient [1300] given by

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2 - 1)}$$

is identical to $\Re$ when $b = 2$. Similarly, Spearman's rank-order correlation coefficient is the measure of association corresponding to Friedman's analysis of variance for ranks test [485], Kendall's coefficient of concordance [739], and Wallis's correlation ratio for ranked data [1411] when $b = 2$ [938]. If the restriction of rank-order statistics is removed and interval measurements are used with $c = 1$, then the permutation version of the Pearson product-moment correlation coefficient [1107] is a special case of $\Re$ when $b = 2$ [938]. In addition, the b-rater extension of the permutation version of Pearson's product-moment correlation coefficient is the measure of association corresponding to a randomized blocks analysis of variance [938].

In 1990 Berry and Mielke provided FORTRAN subroutine AGREE that calculated the generalized measure of agreement $\Re$ and its associated Pearson type III approximate probability value [137]. Subroutine AGREE was constructed for $2 \leq n \leq 20$, $2 \leq b \leq 10$, and $1 \leq c \leq 5$.

5.14.5 Agreement with Two Groups and a Standard

Building on their 1988 article on a generalized measure of agreement, $\Re$, in 1997 Berry and Mielke published two more articles on measures of agreement. In the first article an index of agreement was developed to compare two independent groups of raters [144] and in the second article an index of agreement was developed for the joint agreement between multiple raters and a standard set of responses [146].

Agreement Between Two Independent Groups of Raters

It is often of interest to evaluate the difference between measures of agreement obtained from two independent groups of raters. For example, if written essays are scored by a group of professional educators on a set of criteria and are scored independently by a group of graduate students on the same criteria, it might be of interest to know the difference in agreement between the two groups of raters. To this end, Berry and Mielke developed a test of difference between two independent measures of agreement and provided an associated probability value [144].

Let $\Re_1$ ($\Re_2$) denote the measure of agreement for Group 1 (Group 2) and let μ_1 (μ_2), σ_1^2 (σ_2^2), and γ_1 (γ_2) denote the mean, variance, and skewness, respectively, for Group 1 (Group 2). If $\mathfrak{D} = \Re_1 - \Re_2$, then Berry and Mielke showed that the exact mean, variance, and skewness of $\mathfrak{D}$ under the null hypothesis were given by

$$\mu_{\mathfrak{D}} = 0,$$

$$\sigma_{\mathfrak{D}}^2 = \frac{\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}{\mu_1^2\mu_2^2},$$

and

$$\gamma_{\mathfrak{D}} = \frac{\mu_1^3 \sigma_2^3 \gamma_2 - \mu_2^3 \sigma_1^3 \gamma_1}{\mu_1^3 \mu_2^3 \sigma_{\mathfrak{D}}^3},$$

respectively. Calculation of the exact mean, variance, and skewness of the permutation distribution of $\mathfrak{D}$ permitted utilization of the Pearson type III probability distribution to generate a moment-approximation probability value for an observed value of $\mathfrak{D}$. FORTRAN subroutine DIFFER was provided by the authors in 1997 to test for the difference between two independent groups of raters [146].

Joint Agreement Between Multiple Raters and a Standard

Noting that a number of research problems require the measurement of agreement between multiple raters and a standard (correct) set of responses, in 1997 Berry and Mielke proposed a chance-corrected index that measured the agreement of multiple raters with a standard set of responses [146]. The index was general enough to be used with any level of measurement and with multivariate responses.

If r denotes the number of responses for each of n objects scored by m raters, and the index of the standard set is denoted by s, then the measure of agreement $\Re$ between the m raters and the standard set is given by

$$\Re = 1 - \frac{\delta}{\mu_\delta},$$

where

$$\delta = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[\sum_{k=1}^{r} \left(x_{sjk} - x_{ijk} \right)^2 \right]^{1/2},$$

and μ_δ is the expected value of δ under the null hypothesis. If $m = 1$, $r = 1$, and the responses are categorical, $\Re$ reduces to Cohen's kappa statistic [263]. Berry and Mielke calculated the exact mean, variance, and skewness of the permutation distribution of δ and utilized the Pearson type III probability distribution to generate a moment-approximation probability value for an observed value of $\Re$. FORTRAN subroutine ASAND was provided by the authors in 1997 to test for the difference between multiple raters and a standard set of responses [146].

5.15 Basu and the Fisher Randomization Test

In 1980 Debabrata Basu published a highly controversial article on the Fisher randomization test in *Journal of the American Statistical Association* [86]. The published article was based on an invited talk given at the Southern Regional Education Board (SREB) Summer Research Conference in Statistics at Arkadelphia, Arkansas, on 15 June 1978. Basu examined the Fisher randomization test with respect to sufficiency, common sense, and Fisher's apparent abandonment of randomization tests later in his career. Basu considered permutation tests in comparison to analyses drawing on, for example, normal theory and Bayesian statistics. After comparing the techniques, Basu concluded that permutation tests are deficient because their focus is only on randomization and that that focus disregards important information about controllable and uncontrollable factors. He further wrote that "the Fisher randomization test is not logically viable" [86, p. 575], i.e., the logic of the randomization test procedure is not viable [87, p. 593]. Naturally, this prompted a good deal of debate, most notably by Oscar Kempthorne [723], David Lane [797], Dennis Lindley [828], Donald Rubin [1202], and David Hinkley [627], most of whom defended the Fisher randomization test and disagreed with Basu's contention that the test was illogical. See also a discussion of this debate in a 1994 article by Stephen Senn [1250].

5.16 Still–White and Permutation Analysis of Variance

In 1981 Still and White published an important paper on permutation tests as alternatives to F tests for a variety of analyses of variance designs [1324]. They observed that in experimental psychology it is usually difficult to show (1) that sampled populations meet the normality and homogeneity assumptions for conventional F or t tests, or (2) that sampled populations are similar to those populations sampled in Monte Carlo experiments designed to demonstrate the robustness of conventional F or t tests. They then argued that a test that makes weaker assumptions without sacrificing power or versatility would be preferable. They suggested the use of approximate-randomization (resampling) permutation tests utilizing a Monte Carlo test procedure wherein a random sample of all possible permutations of the observed data was generated and compared with the observed data with respect to a suitable test statistic [1324].

Perhaps anticipating criticism of an approximate randomization test, Still and White noted that the approximate randomization test had few advocates as many researchers believed that different investigators might obtain somewhat different results on the same set of data, even if they used the same subjects, the same test statistic, and the same pseudorandom number generator, due to the use of a different seed; see in this regard, a 1991 article by Spino and Pagano [1308, p. 350].[29] Still and White argued that this was not a valid criticism of Monte Carlo permutation methods and countered with the fact that one may say of any experiment that a different investigator is likely to obtain a different result even if the same subjects are used, as the random assignments of treatments to subjects is likely to be different [1324, p. 246].

Still and White carried out Monte Carlo experiments on four designs: (1) a one-way randomized analysis of variance with three levels and five observations at each

[29] This criticism is largely moot today with fast permutation generators and the selection of 1,000,000 or more random permutations of the observed data being quite common.

level and with varying degrees of separation among the levels; (2) a 2×2 randomized factorial analysis of variance with five observations at each of the four combinations of levels and with varying degrees of separation between the levels; and (3) and (4) the same as (1) and (2) only with blocking introduced to give a repeated-measures design with five subjects. On the basis of these simulated experiments they concluded that, where the sampled population distribution was unknown, the approximate randomization test was preferable to a conventional F test in all four cases [1324].

Six years later, in 1987, Bradbury corrected some errors in the methodology of Still and White [200], observing that the simulations in Still and White were based on the use of an incorrect value of σ_x for non-normal data (1,000 instead of 2,706) and the use of insufficiently large degrees of separation for main and interaction effects for both normal and non-normal data types [200, p. 178]. Bradbury repeated the analyses in Still and White with the errors corrected. Some efficiency was gained by removal of the permutation invariant components of the test statistics, where possible. As in Still and White, four designs were considered: (1) a completely randomized design with five different subjects in each of three treatments, (2) a randomized block design with three treatments and five subjects as blocks, (3) a 2 × 2 factorial design with five different subjects per treatment, and (4) a three-way design given by replicating the 2 × 2 factorial design over five subjects [200, p. 178].

Bradbury concluded, in contrast to Still and White, that approximate randomization tests tended to have slightly higher power for lower levels of effect, whereas conventional F tests tended to have slightly higher power for higher levels of effect. Finally, Bradbury strongly recommended the use of correctly-formulated randomization and approximate randomization tests whenever computational facilities were available [200, p. 187].

5.17 Walters and the Utility of Resampling Methods

In 1981 Walters published an article that demonstrated the usefulness of resampling methods for estimating probability values in two types of permutation tests, with examples: an analysis of $r \times c$ contingency tables and an analysis of randomized blocks analysis of variance [1413]. Noting that an alternative to complete enumeration is to take a random sample from the permutation distribution to provide an estimate of the significance probability, Walters first considered an $r \times c$ contingency table with fixed marginal frequency totals. Utilizing an example analysis of a 2 × 4 contingency table, he computed $\chi^2 = 6.04$ with a probability value of 0.110 based on three degrees of freedom; Walters' 2 × 4 data table is listed in Fig. 5.11. The result of the chi-squared analysis was then compared with an exact permutation test of 804 arrangements of the $n = 40$ observations, yielding an exact probability value of 0.144 and a resampling-approximation probability value of 0.139, based on 5,000 random arrangements of the cell frequencies with fixed marginal frequency totals [1413].

Fig. 5.11 Walters' 2 × 4 contingency table with $n = 40$ observations [1413, p. 292]

8	5	7	3	23
2	5	3	7	17
10	10	10	10	40

Fig. 5.12 Walters' 4 × 3 contingency table with $n = 48$ observations [1413, p. 292]

4	2	1	7
6	5	3	14
1	5	5	11
1	6	9	16
12	18	18	480

Unfortunately, Walters erred in summing the hypergeometric probability values less than or equal to the hypergeometric probability value of the observed contingency table ($p = 0.0018$) instead of summing the probability values associated with the chi-squared values greater than or equal to the observed chi-squared value ($\chi^2 = 6.04$). The correct method had previously been detailed by Radlow and Alf in 1975: order terms by their discrepancies from the null hypothesis instead of by their probability values (q.v. page 249) [1150]. As Walters explained:

> [t]he significance probability in the exact test is the sum of all probabilities less than or equal to [the observed probability value], arising from various configurations of cell frequencies [1413, p. 290].

Walters committed the same error in the resampling-approximation analysis of the example data. Fortunately, the two methods, summing the hypergeometric probability values and summing the hypergeometric probability values associated with the test statistic values, coincidentally yielded the same probability value of 0.144 for the example data analyzed. Walters was less fortunate with his second example analysis based on a 4 × 3 contingency table, which is reproduced in Fig. 5.12. Here he found $\chi^2 = 12.044$ with a probability value of 0.061 based on six degrees of freedom. An exact permutation test of 171,512 arrangements of the $n = 48$ observations yielded an exact probability value of 0.066 and a resampling-approximation probability value of 0.066, based on 5,000 random arrangements of the cell frequencies with fixed marginal frequency totals.[30] These probability values were based on the sum of the hypergeometric probability values less than or equal to the hypergeometric probability value associated with the observed contingency table, i.e., Fisher's exact probability test. The correct exact probability value based on the sum of the hypergeometric probability values associated with the chi-squared values greater than or equal to the observed chi-squared value is 0.059.

In the randomized block analysis, Walters analyzed the data correctly:

> [t]he significance probability is the proportion of permutations returning a test statistic [value] as extreme as the observed [test statistic] value [1413, p. 293].

[30] Walters reported an exact probability value of 0.066, but the correct value should be 0.067 as the exact probability value is actually 0.066628.

For an example analysis of $t = 4$ treatments on $b = 8$ blocks, Walters found a probability value for the conventional F statistic of 0.035 and a resampling probability value of 0.017, based on an undisclosed number of random permutations.[31] Walters then log transformed the original data and reran the randomized block analysis, finding an F test probability value of 0.055 and a resampling probability value of 0.057, ignoring the fact that all permutation tests are distribution-free and, therefore, any transformation of the data was unnecessary.

5.18 Conover–Iman and Rank Transformations

Harking back to the heyday of rank tests in the 1930s and 1940s, in 1981 William Conover and Ronald Iman published an article on rank transformations as a bridge between parametric and non-parametric statistics [273]. The motivation for the article was to increase the visibility and usability of non-parametric techniques. In this brief article of only six pages they called attention to three potential uses of the method of rank transformations: (1) as a pedagogical technique for incorporating non-parametric statistics into introductory courses in statistics, (2) as a method for using existing statistical packages for computing non-parametric statistics, and (3) as a useful tool for developing new non-parametric methods in situations where satisfactory parametric procedures already exist. Comments provided by Michael Fligner and Gottfried Noether in 1981 largely agreed with the suggestions proposed by Conover and Iman [471, 1040]. What is notable about this article and the exchanges is that, at this late date of 1981, no mention was made of permutation tests.

5.19 Green and Randomization Tests

In 1981 Bert Green wrote a lukewarm review of Edgington's newly published book (1980) on *Randomization Tests* that revealed that rank tests were not dead and that permutation (randomization) tests were still not widely accepted [549]. In this review, Green pointed out that, based on his own extensive Monte Carlo studies of robustness, randomization tests were too much like t tests, and shared all the flaws of t tests [549, p. 495]. Green noted that when the t test is non-robust, so is the randomization test, the reason being that both tests use the raw data, rather than some transformation of the data, such as ranks. Thus extreme differences, such as created by outliers, overpower small differences from the mean due to squaring of the differences.

[31] Resampling was required in this case as an exact test would have required generating $(t!)^b = (4!)^8 = 110{,}075{,}314{,}176$ F values.

5.20 Gabriel–Hall and Rerandomization Inference

In 1983 K. Rubin Gabriel and William J. Hall provided computationally-efficient resampling methods for hypothesis testing, estimation via confidence intervals, and power evaluation for a class of experiments without assuming random sampling from a population [489]. Contained within this article was a plea for wider use of permutation methods based on the authors' "conviction that this methodology has a much greater role to play in scientific inference than has been assigned to it in the past" [489, p. 833]. Gabriel and Hall noted in 1978 that Tukey, Brillinger, and Jones had stated:

> [t]he device of judging the strength of evidence offered by an apparent result against the background of the distribution of such results obtained by replacing the actual randomization by randomizations that might have happened seems to us definitely more secure than its presumed competitors, that depend upon specific assumptions about distribution shapes or about independence. . . . On balance, we recommend using a re-randomization analysis . . . [216, p. D1].

In addition, Gabriel and Hall acknowledged their general agreement with Kempthorne's observation in 1955 that "when one considers the whole problem of experimental inference . . . there seems little point in the present state of knowledge in using [a] method of inference other than randomization analysis" (Kempthorne, quoted in Gabriel and Hall) [719, p. 966].

5.21 Pagano–Tritchler and Polynomial-Time Algorithms

In 1983 Pagano and Tritchler noted that the exact permutation test of greatest power is based on all permutations of the observed data [1083]. They argued that the power loss incurred by observing a random sample of all permutations was not great enough to warrant the computational burden of complete enumeration. Consequently, the preferred exact statistical method was usually not employed. Additionally, they observed that an unappealing feature of resampling-approximation permutation tests is the possibility of different researchers obtaining different results with the same data [1083, p. 435]; see also a 1991 paper on this topic by Spino and Pagano [1308, p. 350], and a refutation of this argument by Still and White in 1981 [1324].

To this end, Pagano and Tritchler presented polynomial-time algorithms for finding the permutation distribution of any statistic that was a linear combination of some function of either the original observations or the ranks of the observations [1083, p. 83]. The algorithms required polynomial time as contrasted with complete enumeration, which required exponential time. The savings in time was effected by first calculating and then inverting the characteristic function of the statistic.[32]

[32] For a lucid and cogent description of the Pagano–Tritchler algorithm, see a 1998 article by Gebhard and Schmitz in *Statistical Papers* [503].

5.22 Welch and a Median Permutation Test

In 1987 William Welch proposed a permutation test using the median instead of the mean in matched-pairs designs [1431]. The motivation for this approach was consideration of data sets in which outliers are anticipated. Noting that the sample mean is a poor estimator of location for heavy-tailed distributions, Welch advocated the use of either medians or trimmed means; see also a 1983 paper on this topic by Rosenberger and Gasko [1194]. Employing a randomization procedure, Welch developed resampling probability values and confidence intervals for the sample median and trimmed means.

A comparison of the randomized means test, the randomized median test, and a conventional t test for a matched-pairs example data set with and without a single outlier pair revealed that both the randomized means test and the t test showed "extreme sensitivity" to the exclusion of the outlier pair [1431, p. 613]. Welch concluded that while much of the research in permutation inference had concentrated on test statistics suggested by classical parametric techniques, the permutation approach applied equally well to robust statistics, such as the median and trimmed mean; see also in this regard, a 1985 paper by Diane Lambert in *The Annals of Statistics* [791].

5.23 Boik and the Fisher–Pitman Permutation Test

Also in 1987, Robert Boik published an article on the Fisher–Pitman permutation test in which he investigated the robustness of the test as an alternative to the conventional analysis of variance F test when the variances were heterogeneous [175]. Boik argued that while the permutation test is very attractive as a test of the equality of distributions because it retains its stated test size without any distributional assumptions, it is not as attractive as a test of the equality of location parameters as it retains its stated test size only under equality of all nuisance parameters [175, p. 27].

Boik compared the size of the Fisher–Pitman permutation test of equality of means to the size of the conventional F test in small samples when the variances were unequal. He utilized the first two moments of the beta distribution to approximate the permutation distribution of five sets of three treatments with very small samples with values of $\{1, 1, 7\}$, $\{2, 2, 5\}$, $\{2, 3, 4\}$, $\{2, 3, 5\}$, and $\{2, 4, 5\}$, and also of four sets of four treatments with sample values of $\{2, 2, 2, 10\}$, $\{2, 3, 3, 8\}$, $\{2, 3, 4, 7\}$, and $\{4, 4, 4, 4\}$. Based on computer simulations of 1,000 replications drawn from normally-distributed populations having equal means but unequal variances, he concluded that (1) the normal-theory F test can be very sensitive to variance heterogeneity, and (2) typically, the difference between the normal theory and the permutation test sizes was negligible. Consequently, neither test was found to be robust to variance heterogeneity [175, pp. 36–37].

5.24 Mielke–Yao Empirical Coverage Tests

A class of multi-sample tests that immediately extends to $g \geq 2$ samples is the collection of g-sample empirical coverage tests [965, pp. 335–337]. The g-sample empirical coverage tests described by Mielke and Yao in 1988 and 1990 are direct extensions of goodness-of-fit coverage tests [989, 990]. Following the notation of Mielke and Yao, if $x_1, \ldots, x_n$ is a random sample from an unknown distribution function F, and F_0 is a specified continuous distribution function, the goodness-of-fit coverage test is given by

$$A_v = \sum_{i=1}^{n+1} \left| C_i - (n+1)^{-1} \right|^v ,$$

where $v > 0$, $C_i = F_0(x_{i,n}) - F_0(x_{i-1,n})$ for $i = 1, \ldots, n+1$ are called coverages or spacings, $x_{1,n} < \cdots < x_{n,n}$ are the order statistics of $x_1, \ldots, x_n$, $x_{0,n} = -\infty$, and $x_{n+1,n} = +\infty$. While Major Greenwood introduced A_2 in 1946 [552, 965, pp. 275–277], Maurice Kendall suggested that A_1 should also be considered [732]. A goodness-of-fit coverage test corresponds to a two-sample empirical coverage test in the same manner that the Kolmogorov goodness-of-fit test [766, 965, p. 274] corresponds to the two-sample Kolmogorov–Smirnov test [965, 1283, pp. 334–335]. However, the extension from two samples to g-samples ($g \geq 2$) is straightforward for empirical coverage tests.

As Mielke and Yao described the g-sample empirical coverage tests, let $x_{1|i} < \cdots < x_{n_i|i}$ be the n_i order statistics associated with the ith sample, $i = 1, \ldots, g$,

$$N = \sum_{i=1}^{g} n_i ,$$

$$F_N(x) = \frac{\text{number of observed values among the } N \text{ pooled values} \leq x}{N+1} ,$$

and $F_N(x) = 1$ if x is greater than or equal to the least upper-bound of the domain of x. As defined, $F_N(x)$ differs slightly from the empirical distribution function of the pooled samples. The $n_i + 1$ empirical coverages associated with the n_i observed values of the ith sample are denoted by

$$C_{j|i} = F_N(X_{j|i}) - F_N(j-1|i) \qquad \text{for } j = 1, \ldots, n_i + 1,$$

where $F_N(x_{0|i}) = 0$, $F_N(x_{n_i+1|i}) = 1$, and $x_{0|i}$ and $x_{n_i+1|i}$ are the greatest-lower and least-upper bound values of the unknown population domain of x under the null hypothesis (H_0), respectively. Thus,

$$\sum_{j=1}^{n_i+1} C_{j|i} = 1 \qquad \text{for } i = 1, \ldots, g.$$

For a simple example to illustrate an empirical coverage test, let $g = 2$, $n_1 = 2$, $n_2 = 3$, and $x_{1|2} < x_{2|2} < x_{1|1} < x_{3|2} < x_{2|1}$ be the sample order statistics. Then the empirical coverages for this example are

$$C_{1|1} = 1/2,\ C_{2|1} = C_{3|2} = C_{4|2} = 1/3,\ \text{and}\ C_{3|1} = C_{1|2} = C_{2|2} = 1/6.$$

The null hypothesis specifies that the g samples come from a common continuous distribution and the expected value of $C_{j|i}$ under H_0 is $(n_i + 1)^{-1}$. The g-sample empirical coverage test statistic is then given by

$$B_v = \sum_{i=1}^{g} a\left(\frac{n_i}{N}\right) \sum_{j=1}^{n_i+1} \left| C_{j|i} - (n_i + 1)^{-1} \right|^v,$$

where $v > 0$ and $a(\cdot)$ is a positive weighting function. In 1981 Rao and Murthy proposed a statistic that is equivalent to B_2 with $g = 2$ and $a(\cdot) = 1$ [1156]. Under H_0 and given a condition conjectured to hold in general, but verified only for $v = g = 2$, the distribution of B_v is asymptotically normal as $N \to \infty$ when $v > 1/2$ [989]. Provided the g samples are sufficiently large, the test is able to detect any nontrivial alternative to H_0. Since the total number of equally-likely events is

$$M = \frac{N!}{\displaystyle\prod_{i=1}^{g} n_i!},$$

the exact probability (P) value is given by

$$P = \frac{\text{number of the } M \text{ events in which } B_v \geq B_{vo}}{M},$$

where B_{vo} is the observed value of B_v. When M is large, approximate resampling or Pearson type III probability methods are essential. The exact mean of B_v under H_0 is given by

$$\mu_v = \sum_{i=1}^{g} \left(n_i + 1\right) \sum_{j=1}^{N-n_i+1} \left| j(N + 1)^{-1} - (n_i + 1)^{-1} \right|^v \binom{N-j}{n_i - 1} \bigg/ \binom{N}{n_i},$$

where $v > 0$ and $a(\cdot) = 1$. To obtain a Pearson type III P value given by $P(B_v \geq B_{vo}|H_0)$, the exact variance and skewness of B_v under H_0 (i.e., σ_v^2 and γ_v) are required. However, the calculation of σ_v^2 and γ_v is exceedingly time-consuming, even for moderate sample sizes, due to the multiple looping structures involved. Consequently, a Pearson type III algorithm initially contains μ_v and then, based on L independent simulations of B_v denoted by $B_{v1}, \ldots, B_{vL}$, evaluates estimators of σ_v^2 and γ_v given by

$$\tilde{\sigma}_v^2 = \frac{1}{L}\sum_{i=1}^{L}\left(B_{vi} - \mu_v\right)^2$$

and

$$\tilde{\gamma}_v = \frac{\dfrac{1}{L}\displaystyle\sum_{i=1}^{L}\left(B_{vi} - \mu_v\right)^3}{\tilde{\sigma}_v^3}.$$

Furthermore, a resampling-approximation probability (P) value is given by

$$P = \frac{\text{number of the } L \text{ events in which } B_v \geq B_{vo}}{L}.$$

Simulated moment results suggested that $v = a(\cdot) = 1$ constituted reasonably good choices for B_v [990]. Thus, the probability value of the observed B_1 with $a(\cdot) = 1$ was based on the exact mean of B_1, an estimated variance of B_1 involving 4,000 random simulations, and the normal distribution approximation. Because of the vastly increased speed of modern-day computers, the same probability value would today be estimated using an approximate resampling estimate with perhaps $L = 1{,}000{,}000$ simulations of B_1. The latter resampling estimate of the observed B_1 probability value would avoid the obvious problem associated with the assumption of the normal distribution.

5.25 Randomization in Clinical Trials

In 1988 a series of six articles appeared in the journal *Controlled Clinical Trials* that stemmed from a workshop on randomization for the 1986 annual meeting of the Society for Clinical Trials organized by John Lachin of George Washington University. The six articles, all published in the same issue, consisted of a foreword on "Properties of randomization in clinical trials"by Lachin [785], "Statistical properties of randomization in clinical trials" by Lachin [787], "Properties of simple randomization in clinical trials" by Lachin [786], "Properties of permuted-block randomization in clinical trials" by Matts and Lachin [906], "Properties of the urn randomization in clinical trials" by Wei and Lachin [1423], and "Randomization in clinical trials: Conclusions and recommendations" by Lachin, Matts, and Wei [788]. In addition, a response by Leslie Kalish followed 2 years later in 1990 [706].

Altogether, the 6 articles comprised 88 journal pages and only a brief summary can be attempted here. Differences between a population and a permutation model as bases for statistical tests were reviewed (q.v. page 3), and it was argued that the Neyman–Pearson population model can only be invoked in clinical trials as an untestable assumption, rather than being formally based on sampling at random from a defined population. On the other hand, the authors noted that the Fisher permutation model based on the randomization of treatments to subjects required no assumptions regarding the origin of the samples of patients studied. The large-sample permutation distribution of the family of linear rank tests was described as a basis for easily conducting a variety of permutation tests. Stratified analyses, analyses when some data are missing, and regression model analyses were also discussed. The articles concluded with 15 recommendations regarding the use of permutation tests in controlled clinical trials. See also a 1994 article on “Fisher’s game with the devil” by Stephen Senn in *Statistics in Medicine* [1250].

5.26 The Period from 1990 to 2000

The period in and around 1990–2000 witnessed an explosion of journal articles on permutation methods in a wide variety of disciplines and research areas, e.g., animal behavior [1013], archaeology [970], atmospheric science [959, 1285], biology and biometrics [14, 227, 706, 785–788, 891, 906, 1335, 1338, 1396, 1423], biostatistics [850, 854, 855], chemistry [827, 1395, 1490], clinical trials [99, 411, 1381], dental research [266, 803, 1207, 1208], earth science [943, 1086], ecology [1143, 1293], engineering [16], forest research [1159], geology [541, 1071, 1072], human genetics [1414], medicine and environmental health [170, 235, 648, 1051], pharmacology and physiology [849], psychology and education [141, 950], toxicology and environmental safety [1134], wood science [1114, 1115], and zoology and taxonomy [416–418, 1373].

This period was also characterized by the publication of a number of articles and tutorials that attempted to introduce or promote permutation methods to a variety of audiences, such as psychologists [88, 908], econometricians [748], high-school mathematics teachers [61], chemists [1395], researchers in biomedicine and clinical trials [99, 648, 850, 854–856], and even statisticians [1432].

Earlier undertakings on the development of permutation methods, coupled with high-speed computers and efficient algorithms, provided a solid foundation for permutation methods in the 1990s. While much of the focus in this period was on applying permutation methods to specific research problems, work continued unabated on the development of permutation methods for new research areas and the incorporation of permutation algorithms into various statistical packages [250]. There were so many articles published in this period, with over 100 journal articles appearing each calendar year, that it is not possible to summarize all of them. Thus, it is necessary to carefully select those that are most representative and those with the greatest impact, scope, and importance.

5.27 Algorithms and Programs

In 1991 Oden published an article on the allocation of effort in Monte Carlo simulations of permutation tests in which he determined the optimal choice for the inner (of two) loops in exact permutation tests. In 1992 Kromrey, Chason, and Blair announced the availability of SAS algorithm PERMUT that provided approximate permutation tests for one- and two-sample analyses [773]. Although not explicitly stated, the algorithm apparently utilized resampling-approximation permutation methods rather than exact permutation methods. Also in 1992, Edgington and Khuller published a FORTRAN computer program for trends in repeated-measures (blocked) data [395]. As they explained, the permuting of data for a repeated-measures test of trend with k levels of a treatment and n subjects rearranges the order of the measurements over the k treatment levels for each of the n subjects. For efficiency, the program permuted the data randomly, providing a random sample (resampling) of the $(k!)^n$ possible arrangements of the data.[33] The program produced two probability values: one- and two-tailed. As Edgington and Khuller explained, the one-tailed probability value is the proportion of data arrangements providing Pearson product-moment correlation coefficients as large as or larger than the Pearson product-moment correlation coefficient for the obtained data. The two-tailed probability value is the proportion of data arrangements providing Pearson product-moment correlation coefficients with absolute values as large or larger than the correlation coefficient for the obtained data [395]. Finally in 1992, Ko and Ruskey developed recursive algorithm GENBAG to generate permutations by implementing on both constant amortized time and the interchange property, where constant amortized time is linear in the number of permutations and the interchange property is such that successive permutations differ only by the interchange of two elements [764].

In 1993 Chen and Dunlap contributed to the growing list of papers on permutation tests by providing SAS programs for testing hypotheses using a resampling-approximation permutation test. The article included SAS code listings for testing the equality of two means, testing the significance of a Pearson product-moment correlation coefficient, and testing the equality of more than two means [250]. In 1995 Onghena and May noted some problems with the SAS program of Chen and Dunlap and set about identifying the problems and correcting the SAS code [1063]. The Onghena and May paper is important in that it identified a number of problems with Monte Carlo resampling procedures that had plagued permutation methods since the introduction of high-speed computers.

First, Onghena and May argued that the original statistic must be among the resampled statistics; that is, the probability value can never be smaller than one over the number of resampled values. Chen and Dunlap's program drew 1,000 resampled values and compared each of them to the original value of the statistic. Onghena and May modified the program to draw 999 resampled values, then added the original value to the 999 to provide 1,000 values, guaranteeing that the original value was included [1063]. This was a long-standing controversy in the early days of resampling permutation methods, but is less consequential now that the number of resampled values is usually much greater than 1,000.

[33] Edgington and Khuller cite $(n)^{k!}$ possible arrangements of the data, but this is obviously incorrect.

Second, Onghena and May felt that 1,000 resampled values were not enough to provide sufficient accuracy with Chen and Dunlap's use of $\alpha = 0.01$ and recommended 5,000–10,000 resampled values. Third, Onghena and May argued that the SAS program of Chen and Dunlap was inefficient in that it unnecessarily stored all the resampled values. Fourth, Onghena and May recommended exact permutation methods whenever the number of resampled values was greater than the number of permutations of the original data, i.e., oversampling.

In 1994 Hilton, Mehta, and Patel presented an algorithm for computing exact Smirnov tests in continuous or categorical data with balanced or unbalanced samples [623]. In 1996 Richards and Byrd published FORTRAN subroutine FISHER for computing the exact probability value of the Fisher–Pitman permutation test for two independent samples [1169]. Noting that the Wilcoxon and Mann–Whitney two-sample rank-sum tests are actually the Fisher–Pitman permutation test applied to the ranks of the original observations, they explained that the subroutine generated exact probability values for those tests when ranks were substituted for the raw score measurements. They also observed that the problem of treating ties with the Wilcoxon or Mann–Whitney two-sample rank-sum tests was automatically solved by using the Fisher–Pitman algorithm. Also in 1996, Hayes published an article in which he provided permutation tests for the Macintosh computer [600], and in 1997 Tracey provided a FORTRAN computer program for computing a randomization test of hypothesized order relations [1369].

It is abundantly evident that permutation methods, both exact and resampling-approximation, depend on high-speed computing and also that resampling-approximation permutation methods, where a large random sample of all possible permutations of the data is examined, depend on computer-based uniform pseudorandom number generators (q.v. page 211). In 1996 Yadolah Dodge of the Statistics Group at the University of Neuchâtel, Switzerland, surveyed existing computer-based pseudorandom number generators, noting deficiencies in length of cycles, repeatability, speed, and approximation to a uniform distribution [355]. In addition, he observed that a uniform pseudorandom number generator should produce a distribution that is "normal," explaining that a distribution is considered to be normal in base 10 if all digits $0, 1, \ldots, 9$ appear with equal frequency in its decimal expansion, as well as all blocks of digits of the same length [355, p. 342].

Dodge proposed that the decimal expansion of π was a natural source of a uniform random-number generator, explaining that such a uniform random-number generator lacked cycles of any length, was widely available, had an excellent approximation to the uniform distribution, and was normal, just as all irrational numbers are normal.[34] Dodge concluded that "[t]he probability that π is normal is hence equal to one" and "[a] random sequence formed by the digits of π will satisfy all statistical tests of randomness with probability one" [355, p. 342].

[34] In 1996 π had been calculated to $3 \times 2^{31} = 6{,}442{,}450{,}938$ decimal digits. On 22 October 2011 Alexander Yee and Shigeru Kondo announced that π had been calculated to 10 trillion digits on a dedicated desktop computer; the execution time was 371 days.

In 1998 Gebhard and Schmitz published two articles on permutation methods [502, 503]. In the first article they showed that permutation methods had optimum properties for both continuous and discrete distributions. A variety of examples illustrated permutation tests for the continuous distributions: normal, gamma, exponential, chi-squared, and Weibull; and for the discrete distributions: Poisson, binomial, and negative binomial. In the second article they formulated an efficient computer algorithm for computing the critical region. The algorithm was based on earlier work by Pagano and Tritchler that utilized a fast Fourier transform (q.v. page 338) [502, p. 83]. In 1998 Berry and Mielke developed a FORTRAN program for permutation covariate analyses of residuals based on Euclidean distances [149].

5.28 Page–Brin and Google

No account detailing the development of computing in this period would be complete without mention of Larry Page and Sergey Brin, the Stanford graduate students who co-founded Google, Incorporated in 1998 and which is now the Internet's most-visited Web site.

L. Page

Lawrence (Larry) Page was born in 26 March 1973 in East Lansing, Michigan, where both his father and mother were professors of computer science at Michigan State University. Page's father was a pioneer in computer science and artificial intelligence and his mother taught computer programming. Page earned a B.S. degree in engineering from the University of Michigan in 1995 and entered Stanford University the same year as a graduate student in computer engineering, where he earned his master's degree. It was while Page was at Stanford that he met Sergey Brin, who was assigned to show Page around the computer science department when Page first entered Stanford [217, 1084].

S.M. Brin

Sergey Mikhaylovich Brin was born on 21 August 1973 in Moscow, Russia. His parents immigrated to the United States when Brin was 6 years old. Like Larry Page, Brin's early education was at a Montessori school. In 1990 Brin enrolled at the University of Maryland, where his father was professor of mathematics, studying computer science and earning his B.A. degree in 1993. Brin then entered Stanford University in 1993 as a graduate student, where he earned his master's degree in computer science in 1995 [218].

Google

As a research project at Stanford University, Brin and Page created a search engine that listed results according to the popularity of the pages. They call the search engine Google as a play on the mathematical term "Googol," which is a 1 followed by 100 zeroes. After raising $1 million from family, friends, and other investors, Brin and Page launched Google in 1998 in the garage of a friend, Susan Wojcicki, at 232 Santa Margarita, Menlo Park, California, which they rented for $1,700 a month.[35] Page ran Google as co-president along with Brin until 2001, when they hired Eric Schmidt as Chairman and CEO of Google. In 2004 Google went public, raising $1.67 billion in an initial public offering. Today, Google is the Internet's most visited Web site, employing more than a million servers around the world to process over a billion search requests every day, accessing an index of trillions of Web pages. At the time of this writing, Larry Page is CEO of Google, Eric Schmidt is Executive Chairman, and Sergey Brin is President of Special Projects. Both Page and Brin receive an annual salary of one dollar [217, 540].

5.29 Spino–Pagano and Trimmed/Winsorized Means

In 1991 Spino and Pagano published two articles on the efficient calculation of the permutation distribution for robust two-sample statistics using either trimmed or Winsorized means [1308, 1309].

Trimming and Winsorizing

Given a sample of n observations, trimming involves removing k of the highest and l of the lowest values, then computing the desired statistic on the remaining $n - k - l$ values. For example, consider $n = 9$ observed ordered values $\{3, 9, 12, 14, 14, 15, 19, 23, 37\}$ and let $k = l = 2$. The trimmed sample of $n - k - l = 9 - 2 - 2 = 5$ values would then be $\{12, 14, 14, 15, 19\}$, where the two highest values (23 and 37) and the two lowest values (3 and 9) have been removed.

Winsorizing, on the other hand, involves substituting the k highest values with the $k - 1$ value and substituting the l lowest values with the $l + 1$ value, then computing the desired statistic on the n values. For example, as before,

(continued)

[35] Sergey Brin married Anne Wojcicki, Susan's younger sister, in May 2007.

consider $n = 9$ observed ordered values $\{3, 9, 12, 14, 14, 15, 19, 23, 37\}$ and let $k = l = 2$. The Winsorized sample of $n = 9$ values would then be $\{\underline{12}, \underline{12}, 12, 14, 14, 15, 19, \underline{19}, \underline{19}\}$, where the underlined values are the $l + k = 2 + 2 = 4$ substituted values.[36]

Conventionally, in both trimming and Winsorizing, k is set equal to l. Typically, trimming is used with means and Winsorizing with standard deviations and variances.

In both papers they argued that conventional statistical tests were not sufficiently robust in the presence of violations of distributional assumptions, and that the problem was especially acute with small samples and with the existence of outliers, noting that Hampel, Ronchetti, Rousseeuw, and Stahel had previously pointed out in 1986 that permutation tests based on the sample mean generally have very little power when outliers are present [582].

Spino and Pagano sought to improve the efficiency of permutation tests by using a more robust statistic; viz., the trimmed or Winsorized mean. This was an approach advocated by Lambert for two-sample tests [791] and by Welch and Gutierrez for the trimmed mean [1433]. To this end they utilized an efficient polynomial-time algorithm previously developed by Pagano and Tritchler [1083] that calculated the characteristic function of the data using a recursive difference equation and then inverted the characteristic function using a fast Fourier transform (q.v. page 338). Although Spino and Pagano recommended an algorithm that provided a computationally simple procedure for calculating the permutation distribution of the trimmed or Winsorized mean in small sample research situations, they raised an interesting, but unanswered, question near the end of their first article [1308]. Since the Pagano and Tritchler algorithm utilized Monte Carlo randomization (i.e., resampling) of trimmed or Winsorized means when comparing two samples, the question was: should the trimming or Winsorizing be done to each randomization of the data prior to, or after, dividing the randomized data into the two samples. Thus, the difference between the two randomization scenarios hinges upon the order in which the random assignment mechanism is invoked—before or after the observations have been trimmed or Winsorized.

Specifically, the first scenario involves trimming or Winsorizing the combined samples by removing or replacing k of the largest observations and l of the smallest observations, then randomly dividing the randomized data into the two samples,

[36] Winsorizing, or Winsorization, is named for the physiologist-turned-biostatistician Charles P. Winsor [1380, p. 18]. It was Charles Winsor who convinced John Tukey to convert from mathematics to statistics while both were at Princeton University's Fire Control Research Office in the 1940s [814, p. 194]. As Tukey noted in his foreword to Volume VI of *The Collected Works of John W. Tukey*, "[i]t was Charlie, and the experience of working on the analysis of real data, that converted me to statistics. By the end of late 1945, I was a statistician rather than a topologist ..." [871, p. xlviii].

as suggested by Lambert [791]. The second scenario randomizes the observations into two samples, then trims or Winsorizes each sample separately by removing or replacing k_1 of the largest and l_1 of the smallest observations from Sample 1 and k_2 of the largest and l_2 of the smallest observations from Sample 2, as suggested by Randles and Wolfe [1153].

Incidentally, of interest is that the most extreme case of either trimming or Winsorizing when half of the ordered data on both sides is trimmed or Winsorized yields the median. Thus, this extreme case is associated with the robust form of MRPP (q.v. page 254) when the data and analysis spaces are congruent [938, 939, 941, 943, 959]. Specifically, when trimming with ordered observations where the number of observations, $n > 1$, is an odd number, $(n - 1)/2$ of the observations are eliminated from both the left and right tails, leaving the observation occupying the $(n + 1)/2$ position, counting from either tail, which is the middle observation or the median. Trimming with an even number of ordered observations, $n > 2$, involves eliminating $n/2 - 1$ observations from both the left and right tails, leaving any value in the interval of median values from the $n/2$ to $n/2 + 1$ ordered observations, where the $n/2$ and $n/2 + 1$ ordered values are typically averaged to determine the median. Winsorizing a set of n ordered observations is similar to trimming a set of ordered observations, except that observations from both the left and right tails are not eliminated, but instead converted from their original values to the nearest adjacent value (q.v. page 347).[37]

5.30 May–Hunter and Advantages of Permutation Tests

In 1993 Richard B. May and Michael A. Hunter, two Canadian psychologists, published a short article on "Some advantages of permutation tests," thereby joining the ranks of those promoting the use of permutation methods for testing of hypotheses such as Eugene Edgington, Alvan Feinstein, Bryan Manly, Oscar Kempthorne, and John Tukey [908]. They laid out in an elementary and very readable fashion the rationale and advantages of permutation tests, illustrating permutation methods with a two-sample test for means. Their description of the permutation model is so concise and captures the essence of permutation tests so well, it is worth quoting, in part:

> [a]s early as 1937, Pitman pointed out that the permutation model approaches significance testing in a fashion backwards to the normal model. With the normal model you must first know something about a theoretical parent distribution… and evaluate the data in light of this. The permutation model starts with the data and generates a set of outcomes to which the obtained outcome is compared. The reference distribution, or permutation distribution, is derived from all possible arrangements of the data [908, p. 402].

[37] Trimming has long been advocated by the psychologist Rand Wilcox. His extensive writings on the subject have provided a modern impetus to the procedure of trimming and, to a lesser extent, Winsorizing [1448–1452].

It should be noted that in their comparison of permutation and normal-theory models, they gave the null hypothesis as $H_0{:}\bar{x}_1 = \bar{x}_2$ and $H_0{:}\mu_1 = \mu_2$, respectively, thereby reinforcing the fact that permutation tests are non-parametric and do not necessarily make inferences about populations, only the observed samples.

5.31 Mielke–Berry and Tests for Common Locations

In 1994 Mielke and Berry presented permutation tests for common locations among g samples with unequal variances [951]. As they explained, in completely randomized experimental designs where population variances are equal under the null hypothesis, it is not uncommon to have multiplicative treatment effects that produce unequal variances under the alternative hypothesis. Mielke and Berry presented permutation procedures to test for (1) median location and scale shifts, (2) scale shifts only, and (3) mean location shifts only. In addition, corresponding multivariate extensions were provided.

Consider a completely randomized experiment where $\Omega = \{\omega_1, \ldots, \omega_N\}$ denotes a finite sample of N subjects obtained from some super population and the sample is exhaustively partitioned into g disjoint groups denoted by $S_1, \ldots, S_g$. Let x_I denote a response measurement for subject ω_I $(I = 1, \ldots, N)$ and let $n_k \geq 2$ be the a priori number of subjects randomly assigned to treatment S_k $(k = 1, \ldots, g)$. Also, let

$$\Delta_{I,J} = \left|x_I - x_J\right|^v$$

denote the distance between the univariate response measurements of subjects ω_I and ω_J (x_I and x_J), where $v > 0$. If $v = 1$, then $\Delta_{I,J}$ is the ordinary Euclidean distance between response measurements and if $v = 2$, $\Delta_{I,J}$ is the squared Euclidean distance between response measurements.

For clarification, consider the pairwise sum given by

$$\sum_{I<J} \left|x_I - x_J\right|^v,$$

where $x_1, \ldots, x_N$ are univariate response values and $\sum_{I<J}$ is the sum over I and J such that $1 \leq I < J \leq N$. Let $x_{1,N} \leq \cdots \leq x_{N,N}$ be the order statistics associated with $x_1, \ldots, x_N$. If $v = 1$, then the inequality given by

$$\sum_{I<J} \left|x_I - x_J\right| \leq \left|N - 2i + 1\right|\left|x_{i,N} - \theta\right|$$

holds for all θ, and equality holds if θ is the median of $x_1, \ldots, x_N$. If $v = 2$, then the inequality given by

$$\sum_{I<J} \left(x_I - x_J\right)^2 \leq \sum_{i=1}^{N} \left(x_i - \theta\right)^2$$

holds for all θ, and equality holds if θ is the mean of $x_1, \ldots, x_N$ [951].

Location-shift power comparisons among the parametric Bartlett–Nanda–Pillai trace test, a permutation test with Euclidean commensuration, a permutation test with Hotelling commensuration with $v = 1$, and a permutation test with Hotelling commensuration with $v = 2$ were conducted using alpha levels of 0.05 and 0.01 for five bivariate distributions: normal, uniform, exponential, log-normal, and Cauchy. They found that the Hotelling commensuration permutation test with $v = 2$ and the Bartlett–Nanda–Pillai trace test performed very well for both the bivariate normal and bivariate uniform distributions. For the bivariate exponential, bivariate log-normal, and bivariate Cauchy distributions, the Euclidean commensuration permutation test and the Hotelling commensuration permutation test with $v = 1$ performed much better than either test based on $v = 2$. Finally, the Euclidean commensuration permutation test performed better than the Hotelling commensuration permutation test with $v = 1$ for all distributions.

5.32 Kennedy–Cade and Multiple Regression

In 1996 Peter Kennedy and Brian Cade published an article on permutation tests for multiple regression [749]. In this defining article they compared and evaluated four generic methods of conducting a permutation test in the context of linear multiple regression, conceding that a universally-accepted application of the permutation test procedure for linear regression with a single predictor already existed. Using the classical linear regression model given by

$$y = \mathbf{X}\beta + \mathbf{Z}\theta + \varepsilon,$$

where β and θ are parameter vectors and $\mathbf{X}$ and $\mathbf{Z}$ are corresponding matrices of observations on explanatory variables, they sought to test $\theta = 0$.

The first method they evaluated to test $\theta = 0$ they called the "shuffle **Z**" method. In this method the F statistic for testing $\theta = 0$ is calculated and compared to F statistics produced by shuffling the **Z** variables as a group. This is a method employed by Oja [1052], Collins [269], and Manly [875], although both Welch [1432] and ter Braak [1345] expressed concerns about the method.

The second method they called the "shuffle y" method. In this method, advocated by Manly [875, pp. 91–111], the F statistic for testing $\theta = 0$ is calculated and compared to F statistics produced by shuffling the y variable.

The third method they called the "residualized y" method. They noted that Levin and Robbins [822] and Gail, Tan, and Piantadosi [491] suggested this method whereby they residualized y for **X** and then treated residualized y as the dependent variable in a regression on **Z**. In this method the F statistic for testing $\theta = 0$ is calculated and compared to F statistics produced by shuffling $\mathbf{Z}$ on the residualized y variable.

The fourth method was called the "residualize both y and $\mathbf{Z}$" method. In this method both y and $\mathbf{Z}$ are residualized for $\mathbf{X}$ and then shuffled residualized y is regressed on residualized $\mathbf{Z}$. Kennedy and Cade explained that Beaton [89] and Freedman and Lane [478] suggested regressing y on $\mathbf{X}$, shuffling the residuals from this regression and adding them to the predicted y to form a new y vector which is then regressed on $\mathbf{X}$ and $\mathbf{Z}$. They noted that this method is identical to the fourth method in which both y and $\mathbf{Z}$ are residualized.

The four generic methods of conducting a permutation test in the context of multiple regression were evaluated using Monte Carlo studies based on 1,000 replications. Kennedy and Cade recommended the residualize y and $\mathbf{Z}$ method as it alone had desirable repeated-sample properties [749].

5.33 Blair et al. and Hotelling's T^2 Test

In 1994 Blair, Higgins, Karniski, and Kromrey described multivariate permutation tests that could be substituted for Hotelling's generalized T^2 test [169]. They listed four major limitations of Hotelling's T^2 test:

1. The assumption of population multivariate normality.
2. The need for more subjects than variables.
3. The potential lack of power to detect specific alternatives.
4. The lack of an easily-obtained one-sided testing procedure.

Following the notation of Blair et al., let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$ be p-dimensional vectors denoting observations taken on the ith subject under control and treatment conditions, respectively, let $\mathbf{d}_i = (x_{i1} - y_{i1}, \ldots, x_{ip} - y_{ip})$ denote the p-dimensional difference vector that represents the change in response from control to treatment, and let $-\mathbf{d}_i$ denote the negative vector of $\mathbf{d}_i$.

The significance of statistic t is computed as follows. Let t_{o} denote the value of the test statistic computed on the observed data, and for each of the 2^n possible assignments to the n vectors of $\mathbf{d}_i$, compute the value of the test statistic t. Count the number, $N(t_{\text{o}})$, for which t is equal to or greater than t_{o}, then the exact one-sided probability value of the test is given by

$$p = N(t_{\text{o}})/2^n .$$

Alternatively, define an approximate permutation test as

$$p = N(t_{\text{o}})/M ,$$

where M is the number of resampled random permutations.

Blair et al. then defined three test statistics. The first test statistic, t_{sum}, was defined as

$$t_{\text{sum}} = \sum_{j=1}^{p} t_j ,$$

where t_j denoted the usual one-sample t statistic computed on the jth element of $\mathbf{d}$. The second test statistic, $t_{|\text{sum}|}$, was defined as

$$t_{|\text{sum}|} = \sum_{j=1}^{p} |t_j| .$$

The final test statistic, $t_{\max}$, was defined as

$$t_{\max} = t'_j ,$$

where t'_j was equal to the t_j that was greatest in value. Blair et al. then conducted a Monte Carlo study based on 1,000 random permutations to compare the power of the three test statistics to that of Hotelling's generalized T^2 test under a variety of treatment-effect models [169].

5.34 Mielke–Berry–Neidt and Hotelling's T^2 Test

In 1996 Mielke, Berry, and Neidt published a new permutation procedure for Hotelling's multivariate matched-pairs T^2 test [982]. They explained that since Hotelling's T^2 test obtains a vector of measurements on each subject in each of two time periods, the test is applicable in two different analyses. Consider n subjects and c raters. It is possible to block on the n subjects and examine the multivariate difference among the c raters at the two time periods; alternatively, it is possible to block on the c raters and examine the multivariate difference among the n subjects at the two time periods.

In the first analysis Hotelling's T^2 test statistic is distributed under the null hypothesis (H_0) as an F distribution with c and $n - c$ degrees of freedom in the numerator and denominator, respectively. In the second analysis Hotelling's T^2 test statistic is distributed under H_0 as an F distribution with n and $c - n$ degrees of freedom in the numerator and denominator, respectively. Consequently, one of the two analyses will yield a *df* in the denominator that is less than or equal to zero. Moreover, when $n = c$ neither scenario is possible. Mielke et al. developed a multivariate extension of a univariate permutation test for matched pairs that eliminated the problem and was shown to be more discriminating than the Hotelling T^2 test [982].

Following the notation of Mielke et al., let n subjects be associated with a multivariate pre-treatment and post-treatment matched-pairs permutation test and let $\{x_{11r}, \ldots, x_{c1r}\}$ and $\{x_{12r}, \ldots, x_{c2r}\}$ denote c-dimensional row vectors with elements comprised of the c measurements on the rth subject from the pre- and post-treatments, respectively, where $r = 1, \ldots, n$ [982]. Also let

$$\mathbf{d}_{1r} = \begin{bmatrix} d_{11r} \\ \vdots \\ d_{c1r} \end{bmatrix},$$

where $d_{h1r} = x_{h1r} - x_{h2r}$ for $h = 1, \ldots, c$, denote the c-dimensional column vector of differences between the pre- and post-treatment measurements for the rth of n subjects, and let $\mathbf{d}_{2r} = -\mathbf{d}_{1r}$ denote the c-dimensional origin reflection of $\mathbf{d}_{1r}$ for $r = 1, \ldots, n$. The probability (P) under the null hypothesis of the matched-pairs experiment is $P(\mathbf{d}_{1r}) = P(\mathbf{d}_{2r}) = 0.5$ for $r = 1, \ldots, n$. Now consider the test statistic given by

$$\delta = \binom{n}{2}^{-1} \sum_{r<s} \Delta\left(\mathbf{d}_{1r}, \mathbf{d}_{1s}\right),$$

where

$$\Delta\left(\mathbf{d}_{1r}, \mathbf{d}_{1s}\right) = \left[\left(\mathbf{d}_{1r} - \mathbf{d}_{1s}\right)' \left(\mathbf{d}_{1r} - \mathbf{d}_{1s}\right)\right]^{1/2}$$

is the c-dimensional Euclidean distance between the rth and sth subjects' differences, and the sum $\sum_{r<s}$ is over all r and s such that $1 \le r < s \le n$.

If the c measurements are in different units, then the measurements must be made commensurate, i.e., standardized to a common unit of measurement (q.v. page 301). The replacement of d_{hir} with $d_{hir}^* = d_{hir}/\Phi_h$, where

$$\Phi_h = \sum_{r<s} \left\| d_{h1r} \right| - \left| d_{h1s} \right\|$$

for $h = 1, \ldots, c$, ensures that each measurement makes a similar contribution in the c-dimensional space since

$$\sum_{r<s} \left\| \mathbf{d}_{h1r}^* \right| - \left| \mathbf{d}_{h1s}^* \right\| = 1$$

for $h = 1, \ldots, c$. This commensuration is invariant relative to any permutation under the null hypothesis and is termed Euclidean commensuration (q.v. page 301).

If the observed value of δ is denoted by δ_o, then the exact probability (P) value is given by $P\ (\delta \le \delta_o \,|\, H_0)$, i.e., the proportion of the 2^n possible δ values that are less than or equal to δ_o under H_0. If the observed value of Hotelling's T^2 is denoted by T_o^2, then the analogous exact P value is given by $P\left(T^2 \ge T_o^2 \,|\, H_0\right)$, i.e., the proportion of the 2^n possible T^2 values that are greater than or equal to T_o^2 under H_0.

5.35 Cade–Richards and Tests for LAD Regression

In 1996 Brian Cade and Jon Richards developed a permutation test based on proportionate-reduction-in-sums of absolute deviations when passing from reduced- to full-parameter models for testing hypotheses about least absolute deviation (LAD) estimates of conditional medians in linear regression models [233]. Following the notation of Cade and Richards, in the regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{y}$ is an $n \times 1$ vector of observed responses, β is a $(p+1) \times 1$ vector of unknown regression parameters, $\mathbf{X}$ is an $n \times (p+1)$ matrix of predictors, and ε is an $n \times 1$ vector of random errors, the $(p+1) \times 1$ LAD regression estimate of β, $\mathbf{b}$, minimizes

$$\sum_{i=1}^{n} \Big| y_i - \sum_{j=0}^{p} b_j x_{ij} \Big| .$$

For the test statistic, partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is $n \times (p+1)$ and $\mathbf{X}_2$ is $n \times q$, and partition $\beta = (\beta_1, \beta_2)$, where β_1 is a $(p+1) \times 1$ vector of nuisance parameters and β_2 is a $q \times 1$ vector of parameters tested by the null hypothesis H_0: $\beta_2 = 0$ for the model $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$. The statistic proposed by Cade and Richards for testing this null hypothesis compares the proportionate-reduction-in-sums of absolute deviations between estimates for reduced- and full-parameter models. Thus, the observed test statistic for H_0: $\beta_2 = 0$ is

$$T_{\text{o}} = \frac{\text{SAR} - \text{SAF}}{\text{SAF}} ,$$

where

$$\text{SAR} = \min \sum_{i=1}^{n} \Big| y_i - \sum_{j=0}^{p} b_j x_{ij} \Big|$$

and

$$\text{SAF} = \min \sum_{i=1}^{n} \Big| y_i = \sum_{j=0}^{p+q} b_j x_{ij} \Big| .$$

Large values of the observed statistic T_{o} are evidence against the null hypothesis.

Based on power simulations of 5,000 resamplings, Cade and Richards demonstrated that the permutation test using full-model LAD estimates had greater relative power than normal-theory tests employing least-squares estimates for asymmetric, chi-squared error distributions, and symmetric, double-exponential error distributions for models with one ($n = 35$ and $n = 63$) and three ($n = 63$) independent variables. The power simulations demonstrated the low sensitivity of LAD estimates and permutation tests to outlier contamination and heteroscedasticity that was a linear function of X and increased sensitivity to heteroscedasticity that was a function of X^2 for simple regression models [233]. Three permutation procedures for testing partial models in multiple regression were compared by Cade and Richards: (1) permuting residuals from the reduced model, (2) permuting residuals from the full model, and (3) permuting the dependent variable. They found that permuting residuals from the reduced model maintained nominal error rates best under the null hypothesis for all error distributions and for correlated and uncorrelated independent variables [233].

5.36 Walker–Loftis–Mielke and Spatial Dependence

In 1997 Walker, Loftis, and Mielke developed a class of multivariate permutation procedures for spatial dependence, Multivariate Sequential Permutation Analyses (MSPA), and applied the procedures to test for correlation in a geostatistical analysis [1408]; see also a 1991 article by Mielke on this topic [943]. As with most permutation tests, given a finite set of objects an observed spatial pattern of objects is compared against all possible permutations of spatial patterns to determine the exact probability value of the observed spatial pattern. As such, MSPA is closely related to the multi-response permutation procedures (MRPP) discussed in Chap. 4 (qq.v. pages 254–265).

Following the notation of Walker et al., consider a finite set of N objects $\{\omega_1, \ldots, \omega_N\}$ where $\{x_{1I}, \ldots, x_{rI}\}$ denotes r response measurements on object ω_I, $I = 1, \ldots, N$. If the sequence of the objects denotes the observed sequence, then the null hypothesis of MSPA dictates that any one of the $N!$ possible realizations of the observed sequence occurs with equal chance. Thus, under the null hypothesis, the probability of the observed sequence is $1/N!$. The MSPA test statistic is given by

$$\delta = \frac{1}{N-1} \sum_{I=2}^{N} \Delta_{I-1,\,I}\,,$$

where

$$\Delta_{I,\,J} = \left[\sum_{h=1}^{r} (x_{hI} - x_{hJ})^2 \right]^{v/2}$$

is a distance measure between adjacent objects ω_I and ω_J. As with MRPP, if $v = 1$, then $\Delta_{I,J}$ is an r-dimensional Euclidean distance, whereas values of $\Delta_{I,J}$ for $v > 0$ and $v \neq 1$ represent complex distance functions, where the data and analysis spaces are not congruent.

If the observed value of δ, δ_{o}, is small relative to the $N!$ possible values of δ, then a first-order autoregressive spatial pattern is suggested. The exact probability (P) value associated with the observed sequence is given by

$$P = \frac{\text{number of } \delta \text{ values} \leq \delta_{\text{o}}}{N!}.$$

For situations when $N!$ is very large, Walker et al. presented procedures for obtaining resampling-approximation probability values and also for fitting a standardized statistic to the Pearson type III distribution (q.v. page 261).

5.37 Frick on Process-Based Testing

In 1998 Frick published an article that challenged the standard textbook treatment of conventional statistical tests based on random sampling from an infinite population [482]. He termed this treatment the "population-based" interpretation of statistical testing (q.v. page 3), and noted three problems with the population-based treatment:

1. Researchers rarely make any attempt to randomly sample from a defined population.[38]
2. Even if random sampling actually occurred, conventional statistical tests do not precisely describe the population.
3. Researchers do not generally use statistical testing to generalize to a population.

Against the population-based interpretation Frick proposed what he called a "process-based" interpretation, arguing that random sampling is a process, not the outcome of a process; in this regard, see also a 1992 article by Sohn [1292]. To this end, Frick recommended consideration of permutation methods.

To illustrate the process interpretation, Frick explained that R.A. Fisher, in discussing his experiment of the lady tasting tea (q.v. page 58) in which the lady claimed that she could determine which ingredient (tea or milk) was added first, wrote that "the judgments given are in no way influenced by the order in which the ingredients have been added" [461, pp. 15–16]. This, Frick explained, was a claim about process, not populations.

5.38 Ludbrook–Dudley and Biomedical Research

In 1998 John Ludbrook and Hugh Dudley published an influential article titled "Why permutation tests are superior to t and F tests in biomedical research" [856]. The article, appearing as it did in *The American Statistician* attracted a great deal of attention and elicited comments by Douglas Langbehn [798], Vance Berger [98], James Higgins [614], and Colin Mallows [872], as well as a rejoinder by

[38] This reiterated the position held, for example, by Altman and Bland [15], Bradbury [200], Edgington [389], Feinstein [421], LaFleur and Greevy [789], Ludbrook [850], Ludbrook and Dudley [856], and Still and White [1324], that assuming a random sample from an infinite population was untenable in many disciplines.

Ludbrook and Dudley [857]. Ludbrook and Dudley attempted, in this review article, to draw attention to a serious misunderstanding between statisticians and biomedical scientists.[39] They noted that statisticians believe that biomedical researchers conduct most experiments by taking random samples and therefore recommend statistical procedures that are valid under the population model of inference (q.v. page 3). Given that biomedical researchers do not usually employ random sampling, but instead rely on randomization of a nonrandom sample, Ludbrook and Dudley argued that the population model did not apply and strongly recommended statistical procedures based on data-dependent permutations of the observations.

Contained within this article are concise, but thorough, synopses of the two models of statistical inference: the Neyman–Pearson population model and the Fisher permutation model (q.v. page 3), followed by a comparison of the two models illustrated with analyses of the differences between the means of two independent samples. They concluded the article with a quote from Oscar Kempthorne that they felt summarized the substance of the review:

> [w]hen one considers the whole problem of statistical inference, that is of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using [a] method of inference other than randomization analysis (Kempthorne, quoted in Ludbrook and Dudley) [719, p. 966].

5.39 The Fisher Z Transformation

Chapter 5 concludes with an illustration of the utility of permutation methods in a revealing application that could only be accomplished through the use of Monte Carlo (resampling-approximation) permutation methods. In 2000 Berry and Mielke utilized Monte Carlo permutation methods to investigate the Fisher Z transformation of the sample Pearson bivariate product-moment correlation coefficient between variables x and y, r_{xy} [156]. They also investigated two related techniques introduced by Gayen in 1951 [499] and Jeyaratnam in 1992 [685]. In 1915 and 1921 R.A. Fisher obtained the basic distribution of r_{xy} and showed that, when bivariate normality is assumed, a logarithmic transformation of r_{xy},

$$Z = \frac{1}{2} \ln \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right) = \tanh^{-1}(r_{xy}),$$

becomes normally distributed with a mean of approximately

$$\frac{1}{2} \ln \left(\frac{1 + \rho_{xy}}{1 - \rho_{xy}} \right) = \tanh^{-1}(\rho_{xy})$$

[39] John Ludbrook is Professional Research Fellow in the University of Melbourne, Department of Surgery, Royal Melbourne Hospital and Hugh Dudley is Professor Emeritus at the University of London, Department of Surgery, St. Mary's Hospital Medical School.

and a standard error that approaches

$$\frac{1}{\sqrt{n-3}}$$

as n becomes increasingly large [442, 444].

The 1915 paper by R.A. Fisher on “Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population,” published in *Metron*, contained the formula for Z, but it appears almost as an afterthought in the last sentence of the 15 page paper [442, p. 521]. The 1921 paper “On the ‘probable’ error of a coefficient of correlation deduced from a small sample,” published in *Biometrika*, was the second of three papers dealing with the sampling errors of correlation coefficients, and is more extensive than the first [444].[40] The third paper in the series on “The distribution of the partial correlation coefficient” dealt only with partial correlation coefficients and was published in *Metron* in 1924 [447].

Berry and Mielke utilized Monte Carlo permutation methods to compare combinations of sample sizes and population parameters for seven bivariate distributions. Both confidence intervals and hypothesis testing were examined for robustness to non-normality [156]. The seven distributions utilized for the Monte Carlo simulations were the normal distribution, $N(0,1)$, given by

$$f(x) = (2\pi)^{-1/2} \exp(-x^2/2) \; ;$$

the generalized logistic distribution, GL, given by

$$f(x) = (\exp(\theta x)/\theta)^{1/\theta} \left(1 + \exp(\theta x)/\theta\right)^{-(\theta+1)/\theta}$$

with $\theta = 1.0$, 0.1, and 0.01; and the symmetric kappa distribution, SK, given by

$$f(x) = 0.5\lambda^{-1/\lambda} \left(1 + |x|^{\lambda}/\lambda\right)^{-(\lambda+1)/\lambda}$$

with $\lambda = 2$, 3, and 25. The seven distributions ranged from the normal distribution, $N(0,1)$, the logistic distribution, $GL(1.0)$, positively skewed distributions, $GL(0.1)$ and $GL(0.01)$, heavy-tailed distributions, $SK(2)$ and $SK(3)$, to a uniform-like distribution with light tails, $SK(25)$.

For the analyses of confidence intervals, the Monte Carlo analyses were based on $L = 1{,}000{,}000$ random samples of size n generating simulated probability values for the seven bivariate distributions, $N(0,1)$, $GL(1.0)$, $GL(0.1)$, $GL(0.01)$, $SK(2)$, $SK(3)$, and $SK(25)$, with nominal values of $1 - \alpha = 0.90$, 0.95, and 0.99,

[40] It should be noted that the second paper was written in response to a stinging criticism of the 1915 paper by H.E. Soper, A.W. Young, B.M. Cave, A. Lee, and K. Pearson that had appeared in *Biometrika* in 1916 and was titled “On the distribution of the correlation coefficient in small samples” [1297].

population parameter $\rho_{xy} = 0.0$, 0.4, 0.6, and 0.8, and sample size $n = 10$, 20, 40, and 80.[41] A common seed ensured valid comparisons. Two confidence interval analyses were conducted. The first considered confidence intervals based on the Fisher Z transformation and the second considered confidence intervals based on an alternative method proposed by Jeyaratnam in 1992 [685].

For the analyses of hypothesis testing, the Monte Carlo tests of hypotheses were based on the same seven distributions, $N(0, 1)$, $GL(1)$, $GL(0.1)$, $GL(0.01)$, $SK(2)$, $SK(3)$, and $SK(25)$. Each simulation was based on $L = 1{,}000{,}000$ bivariate random samples of sizes $n = 20$ and 80, for $\rho_{xy} = 0.0$ and 0.6, and compared to nominal upper-tail probability values of $\alpha = 0.99$, 0.90, 0.75, 0.50, 0.25, 0.10, and 0.01. A common seed ensured valid comparisons. Two tests of $\rho_{xy} \neq 0.0$ were conducted. The first test of hypothesis was based on the Fisher Z transformation and the second test of hypothesis was based on a transformation proposed by Gayen in 1951 [499].

Based on extensive Monte Carlo simulations, Berry and Mielke concluded that considerable caution should be exercised when using the Fisher Z transform, or related techniques such as those proposed by Gayen and Jeyaratnam, as these methods clearly are not robust to deviations from normality when $|\rho_{xy}| \neq 0.0$ [156, p. 1113]. Most surprisingly, for the heavy-tailed distributions, $SK(2)$ and $SK(3)$, and the skewed distributions, $GL(0.1)$ and $GL(0.01)$, small samples, e.g., $n = 10$, provided better estimates than large samples, e.g., $n = 80$ [156, p. 1112]. The authors explained that larger samples obviously have a greater chance of selecting extreme values than small samples. Consequently, the Monte Carlo containment probabilities became worse with increasing sample size when heavy-tailed distributions were encountered [156, p. 1113].[42]

Fisher originally stipulated that the Z transform was appropriate only when either $\rho_{xy} = 0.0$ or the underlying population distribution was bivariate normal, a requirement that has been consistently ignored by contemporary researchers. Berry and Mielke confirmed that Fisher's statement was absolutely correct [156]. Table 5.4 contains upper-tail Monte Carlo probability values based on $L = 1{,}000{,}000$ for the bivariate $N(0, 1)$ distribution with specified nominal values of $P = 0.99$, 0.90, 0.10, and 0.01, $\rho_{xy} = 0.0$ and 0.6, and $n = 20$ and 80 for the Fisher Z transformation. Inspection of Table 5.4 confirms the close agreement between the probability values

[41] As a testament to computing power in 2000, the authors computed three different confidence intervals at three confidence levels on four values of the population parameter 1,000,000 times with four different sample sizes of 10, 20, 40, and 80; a feat that was inconceivable just a decade earlier.

[42] Authors' note: one of the authors often advises his students regarding the Fisher Z transform:

1. Do not use the Fisher Z transformation.
2. If you do use it, don't believe it.
3. If you do believe it, don't publish it.
4. If you do publish it, don't be the first author.

Adapted from a description of a tiltmeter in *Volcano Cowboys* by Dick Thompson [1357, p. 258].

Table 5.4 Upper-tail probability values compared with nominal values (P) for a bivariate $N(0, 1)$ distribution with Fisher's Z transform on tests of hypotheses for $\rho_{xy} = 0.0$ and $\rho_{xy} = 0.6$ with $n = 20$ and $n = 80$

	$n = 20$		$n = 80$	
P	$\rho_{xy} = 0.0$	$\rho_{xy} = 0.60$	$\rho_{xy} = 0.0$	$\rho_{xy} = 0.6$
0.99	0.9894	0.9915	0.9898	0.9908
0.90	0.9016	0.9147	0.9009	0.9065
0.10	0.0983	0.1098	0.0999	0.1054
0.01	0.0108	0.0126	0.0102	0.0110

Note: Table 5.4 adapted from Berry and Mielke [156, p. 1108]

based on the Fisher Z transform and the nominal values for both $n = 20$ and $n = 80$.

5.40 Looking Ahead

Although the chronicle of the development of permutation methods in this volume concludes with the year 2000, the authors would be remiss not to mention some significant developments after 2000. While permutation methods may be said to have "arrived" in the period from 1980 to 2000, they may be said to have "erupted" in the next decade. Advances in computing, including increased speed, enlarged memory and capacity, canned statistical packages that included permutation add-ons or modules, and the development of a new computer language, R, enabled a virtual explosion of new permutation methods and applications.

After 2000, permutation methods continued to be introduced into, spread to, or expanded in a number of different fields and disciplines, most notably in medicine, psychology, clinical trials, biology, ecology, environmental science, earth science, and atmospheric science. Along with a proliferation of journal articles, a multitude of books on permutation methods appeared. Having all the information collected and organized in one compact source instead of scattered among the many journals in myriad disciplines, made it easier for the user to learn about new and existing permutation methods. Included among these books were volumes by Good [525, 526] in 2000; Good [527], Mielke and Berry [961], and Pesarin [1120] in 2001; Lahiri [790] in 2003; Good [531] in 2005; Good [532] and Hirji [629] in 2006; Edgington and Onghena [396], Manly [877], and Mielke and Berry [965] in 2007; and Pesarin and Salmaso [1122] in 2010.

By 2000, a number of the leading statistics programs in the United States had incorporated permutation methods into their curricula. Many of the permutation methods courses are taught at the graduate level, but the relative simplicity of the permutation approach makes it amenable to students who have little background in statistics or probability theory. Consider that among the top twenty statistics programs in the United States, as identified by the National Research Council (NRC) in 2000, six programs had at least one course that was devoted to permutation methods (Harvard University; the University of Wisconsin, Madison; Texas A&M University; the University of Washington; Stanford University; and the University of California, Los Angeles) and seven programs had at least one course that included permutation statistical approaches as part of the course (Cornell University; University of Chicago; the University of North Carolina, Chapel Hill; Iowa State University; the Pennsylvania State University; Rutgers University; and the University of Washington). Finally, some university-level textbooks in statistics have included sections or chapters on permutation methods, e.g., Howell [656, Chap. 18]. Statistical areas of special interest after 2000 included multiple regression, analysis of variance, measures of agreement and concordance with both linear and quadratic weighting, discriminant analysis, matched pairs, survival analysis, ridit analysis, analysis of trend, robustness and outliers, and multi-way contingency tables.

Three features of permutation methods were especially prominent in the period after 2000. The first entailed an increasing criticism of rank-order statistical procedures with their attendant loss of information due to the substitution of rank-order statistics for numerical values. In lieu of rank-order statistical procedures, many researchers advocated the use of permutation methods that utilized the original numerical values and did not depend on an assumption of normality. The second feature was a criticism of permutation methods based on squared Euclidean distances that gave artificial weight to extreme scores and implied a geometry of the analysis space that differed from the geometry of the ordinary Euclidean data space in question. An alternative was to develop permutation tests and measures based on ordinary Euclidean distances that proved to be very robust relative to outliers, extreme values, and highly skewed distributions. The third feature in this period was a heavy reliance on resampling-approximation permutation methods instead of approximate permutation methods based on the exact first three moments of a continuous distribution that approximated the underlying discrete permutation distribution (i.e., mean, variance, and skewness). Resampling with a large number of replications yielded results arbitrarily close to exact results; moreover, in many cases resampling proved to be more efficient, especially in the analysis of contingency tables.

Beyond 2000

6

This chapter is included simply to document significant contributions to permutation statistical methods following calendar year 2000. By 2001, permutation methods had come of age and advances were comprised more of applications and expansion into new fields and disciplines than the development of new permutation methods that had characterized earlier years, although there were some notable exceptions. While articles on statistical permutation methods continued to be published in the usual fields of statistics, medicine, psychology, public health, environmental science, biology, economics, ecology, and atmospheric science, permutation methods branched out after 2000 into journals in animal research, bioinformatics, business, chemistry, clinical trials, industrial engineering, management, operations research, physiology, and veterinary medicine.

6.1 Overview of This Chapter

By 2001 computing was sufficiently powerful, fast, and available, that permutation methods, for the first time, were both feasible and practical. Moreover, many readily-available statistical packages such as StatXact, SPSS, and Stata had incorporated modules designed to execute a variety of permutation tests. Between 2001 and 2010 more than a dozen books were published on permutation statistical methods, including both exact and resampling-approximation methods. In addition, several 1000 articles were published on permutation methods in a broad array of disciplines. Also by 2001, problems with non-normality were widely recognized and new techniques based on Euclidean-distance measures were introduced and promoted to counter the deleterious effects of outliers and heavy-tailed distributions [978]. In many cases the solution to non-normality of substituting ranks for numerical values was eclipsed by the use of permutation methods that retained the original numerical values and, like rank tests, did not assume normality.

Because of the vast number of articles and books on permutation statistical methods and the scope of interests in this period, it is not possible to do justice

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9_6,
© Springer International Publishing Switzerland 2014

to all the literature. Among the areas of classical statistics explored in this chapter are the analysis of variance, including one-way, factorial, blocked, and cross-over designs, as well as multiple comparison of means analyses; linear regression and correlation, both simple bivariate and multiple, plus quantile, canonical, and tetrachoric correlation; clinical trials analysis; measures of concordance and agreement, including weighted and unweighted measures, and extensions to multiple judges; rank tests, ridit analysis, and power; and Bayesian hierarchical analysis. However, the permutation statistical literature after 2000 was clearly dominated by four interest areas: multiple linear regression, the analysis of variance, measures of agreement and concordance, and contingency table analysis.

This chapter begins with a description of computing after 2000 and concludes with two views of the literature in this period. The first view provides a brief description of representative articles organized by year of publication and the second view provides a more detailed description of selected articles with greater import.

6.2 Computing After Year 2000

One has only to observe the hordes of the digitally distracted trying to navigate a crowded sidewalk with their various smart-phones, pads, pods, and tablets to realize that computing power, speed, and accessibility have finally arrived. As Martin Hilbert documented, in 1986 just 1 % of the world's capacity to store information was in digital format, but by year 2000 digital represented 25 % of the total world's memory [619]. The year 2002 marked the start of the digital age, as 2002 was the year that humankind first stored more information in digital than in analog form. By 2007 over 97 % of the world's storage capacity was digital [619, p. 9]. Moreover, it was estimated in 2012 that 90 % of the data stored in the world had been created in just the previous 2 years. Prior to 2001, data storage was measured in bytes, kilobytes (10^3), and occasionally in megabytes (10^6); now data storage is measured in gigabytes (10^9), terabytes (10^{12}), petabytes (10^{15}), exabytes (10^{18}), zettabytes (10^{21}), and even yottabytes (10^{24}).

In 2000, the Intel Pentium processor contained 42 million transistors and ran at 1.5 GHz. In the spring of 2010, Intel released the Itanium processor, code-named Tukwila after a town in Washington, containing 1.4 billion transistors and running at 2.53 GHz. On 4 June 2013 Intel announced the Haswell processor, named after a small town of 65 people in southeastern Colorado with 1.4 billion 3-D chips and running at 3.50 GHz [1403]. While not widely available to researchers, by 2010 mainframe computers were measuring computing speeds in teraflops. To emphasize the progress of computing, in 1951 the Remington Rand Corporation introduced the UNIVAC computer running at 1,905 flops, which with ten mercury delay line memory tanks could store 20,000 bytes of information; in 2008 the IBM Corporation supercomputer, code-named Roadrunner, reached a sustained performance of one petaflops[1]; in 2010 the Cray Jaguar was named the world's fastest computer performing at a sustained speed of 1.75 petaflops with 360 terabytes of memory; and in November of 2010 China exceeded the computing speed of the Cray Jaguar by 57 % with the introduction of China's Tianhe-1A supercomputer performing at 2.67 petaflops [861].

In October of 2011, China broke the petaflops barrier again with the introduction of the Sunway Bluelight MPP [62]. In late 2011 the IBM Yellowstone supercomputer was installed at the National Center for Atmospheric Research (NCAR) Wyoming Supercomputer Center in Cheyenne, Wyoming. After months of testing, the Wyoming Supercomputer Center officially opened on Monday, 15 October 2012. Yellowstone was a 1.6 petaflops machine with 149.2 terabytes of memory and 74,592 processor cores and replaced an IBM Bluefire supercomputer installed in 2008 that had a peak speed of 76 teraflops. Also in late 2011, IBM unveiled the Blue Gene\P and \Q supercomputing processing systems that can achieve 20 petaflops. At the same time, IBM filed a patent for a massive supercomputing system capable of 107 petaflops.

From a more general perspective, in 1977 the Tandy Corporation released the TRS-80, the first fully assembled personal computer, distributed through Radio Shack stores. The TRS-80 had 4MB of RAM and ran at 1.78 MHz. By way of comparison, in 2010 the Apple iPhone had 131,072 times the memory of the TRS-80 and was about 2,000 times faster, running at one GHz. In 2012, Sequoia, an IBM Blue Gene/Q supercomputer was installed at Lawrence Livermore National Laboratory (LLNL) in Livermore, California. In June of 2012 Sequoia officially became the most powerful supercomputer in the world. Sequoia is capable of 16.32 petaflops—more than 16 quadrillion calculations a second—which is 55 % faster than Japan's K supercomputer, ranked number 2, and more than five times faster than China's Tianhe-1A, which was the fastest supercomputer in the world in 2010.

MareNostrum

To document the rapid advancement of computing speed in the twenty-first century, consider the MareNostrum (Latin for "our sea") supercomputer in the Barcelona Supercomputing Center, the second most powerful computer in Spain and one of the seven supercomputers of the Spanish Supercomputing Network (SSN). The MareNostrum supercomputer is ensconced in the deconsecrated Chapel Torre Girona at the Polytechnic University of Catalonia in Barcelona. MareNostrum weighs 44 tons, has 10,240 central processing units, 20 terabytes of RAM, 280 terabytes of disk storage, and runs at a peak performance speed of 94.21 teraflops while working on models of climate

(continued)

[1] One petaflops indicates a quadrillion operations per second, or a 1 with 15 zeroes following it.

change and other scientific projects. At the time of its construction in 2004 it was considered the fourth most powerful computer in the world, but a short 8 years later its speed had been eclipsed by over 100 new supercomputers and it then ranked as only the 118th most powerful computer in the world in 2012 [1299, 1365].

The future of high-speed computing looks to be very promising for permutation statistical methods. Looking ahead, computer engineers have set their sights extremely high and are designing machines that work in exoscale, i.e., three orders of magnitude above the current frontier.[2] In 2011 Richard Murphy, computer architect at Sandia National Laboratories (SNL) in Albuquerque, New Mexico, headed up a team to produce an energy efficient computer for the Defense Advanced Research Projects Agency (DARPA) called X-caliber [1282]. The approach was based on a distributed architecture where multiprocessors had dedicated sets of memory chips. At the same time, Intel's new project, called Runnemede, uses an innovative technique whereby power is selectively turned on and off to individual components; graphics chip maker NVIDIA leads a research team called Echelon in which the graphics chips execute simultaneous operations, rather than just one operation at a time; and the Angstrom project based at the Massachusetts Institute of Technology is creating a computer that optimizes settings, such as the number of processors in use [1282]. Today's desktop computers rival the supercomputers of the late 1980s and, given the pace of innovation, it is predicted that by 2020, laptops will outperform China's Tianhe-1A supercomputer that presently performs at 2.67 quadrillion operations per second [1282].

In early 2012 two results were announced that promise to bring quantum computers closer to reality. The world's thinnest silicon wire, just one atom high and four atoms wide, was created by a team of researchers from the University of New South Wales, the University of Melbourne, and Purdue University. The silicon wire has the same current-carrying capacity as a copper wire. Michelle Simmons, director of the Centre of Excellence for Quantum Computation and Communication Technology at the University of New South Wales and the project's principal investigator, asserted that the goal of the research was to develop future quantum computers in which single atoms are used for computation. "We are on the threshold of making transistors out of individual atoms," Simmons said [1420]. Indeed, just 6 weeks later in February 2012 another breakthrough in quantum computing was made when a team based at the University of New South Wales, the Korea Institute of Science and Technology, the University of Sydney, and the University of Melbourne announced that a single-atom transistor had been placed by positioning a phosphorus atom between metallic electrodes, also made of phosphorus, on a silicon surface [488].

[2] One exoflops indicates a quintillion floating operations per second, or a 1 with 18 zeroes after it.

On 23 April 2012 Intel unveiled new core processors, code-named Ivy Bridge. These new generation chips for personal computers and hand-held devices were the first to be made with a three-dimensional (3-D) structure, permitting Intel to pack more components into the same space as a two-dimensional (2-D) structure. Based on Intel's 22-nm tri-gate manufacturing process, the new Intel central processing unit (CPU) contained 1.4 billion transistors in a scant 160 mm^2 area (0.25 in.^2). On 4 June 2013 Intel announced a new 4th generation desktop processor code-named Haswell, after a small town in Colorado, with 1.4 billion transistors, 6 MB of cache, and running at 3.50 GHz. The Haswell processor was also based on Intel's 22-nm, tri-gate manufacturing process and incorporated, for the first time, a voltage regulator inside the chip. The Haswell processor provided 50 % longer battery life, three times the amount of standby battery life, and 15 % improved performance when compared to Ivy Bridge. To put this into perspective, the transistors are so small that 100 million of them would fit on the head of a pin, whereas the original transistor built by Bell Laboratories in 1947 was large enough to be pieced together by hand [244, 1151, 1253].

By 2010, computing power was finally sufficient to accommodate the needs of computational statisticians utilizing permutation tests. Keller-McNulty and Higgins concluded on the basis of Monte Carlo results that there was little reason to conduct exact permutation tests, recommending that researchers use only 1,600 random samples [714]. Bailer [48], Kim, Nelson, and Startz [756], and McQueen [917] used only 1,000 random permutations in their studies, and Edgington [391] in 1969 claimed that 999 random permutations of the data (plus the original data arrangement) were sufficient. Dwass [368] in 1967 argued that 10,000 random permutations provided results nearly as powerful as complete enumeration, and Edgington and Khuller [395] concurred. Manly [875, pp. 32–36] and Noreen [1041, p. 15] argued that for testing at the 0.05 level of significance, 1,000 random permutations was sufficient. On the other hand, Fitzmaurice, Lipsitz, and Ibrahim concluded that 200 permutations resulted in a test with a correct type I error rate [464, p. 944]. Because of increasing computing power, however, by 2010 probability values based on exact enumeration sometimes exceeded 10,000,000 permutations and resampling probability values based on 1,000,000 random permutations were not only recommended [696], but common [965].

Increased computational efficiency paved the way for the introduction of a number of software statistical packagesfor permutation tests, now widely available to computational statisticians. Among the most available and popular software packages for permutation tests are Box Sampler (Microsoft Corp., Redmond, Washington), S-PLUS (MathSoft, Inc., Seattle, Washington), Statistica (StatSoft, Inc., Tulsa, Oklahoma), SPSS (SPSS, Inc., Chicago, Illinois), SAS (SAS Institute, Inc., Cary, North Carolina), Stata (StataCorp LP, College Station, Texas), Blossom Statistical Software (Fort Collins Ecological Science Center, Fort Collins, Colorado), Resampling Stats (Resampling Stats, Inc., Arlington, Virginia), Statistical Calculator (StatPac, Bloomington, Minnesota), StatXact (Cytel Software Corp., Cambridge, Massachusetts), Systat (Systat Software, Inc., Chicago, Illinois), and Testimate (Institute for Data Analysis and Study Planning, Munich, Germany).

6.3 Books on Permutation Methods

In addition to permutation statistical software, the period after 2000 saw the publication of a number of books on permutation methods, including volumes on *Data Analysis by Resampling: Concepts and Applications* by Clifford Lunneborg in 2000 [858]; a second edition of *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* by Phillip Good in 2000 [526]; a second edition of *Permutation, Parametric and Bootstrap Tests of Hypotheses* by Phillip Good in 2000 [525]; a second edition of *Resampling Methods: A Practical Guide to Data Analysis* by Phillip Good in 2000 [527]; *Permutation Methods: A Distance Function Approach* by Paul Mielke and Kenneth Berry in 2001 [961]; *Multivariate Permutation Tests: With Applications in Biostatistics* by Fortunato Pesarin in 2001 [1120]; *Resampling Methods for Dependent Data* by Soumendra Lahiri in 2003 [790]; a third edition of *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, retitled *Permutation, Parametric and Bootstrap Tests of Hypotheses* to focus more on parametric and bootstrap procedures by Phillip Good in 2005 [531]; a third edition of *Resampling Methods: A Practical Guide to Data Analysis* by Phillip Good in 2006 [532]; *Exact Analysis of Discrete Data* by Karim Hirji in 2006 [629]; a fourth edition of *Randomization Tests* by Eugene Edgington and Patrick Onghena in 2007 [396]; a third edition of *Randomization, Bootstrap and Monte Carlo Methods in Biology* by Bryan Manly in 2007 [877]; a second edition of *Permutation Methods: A Distance Function Approach* by Paul Mielke and Kenneth Berry in 2007 [965]; a second edition of *Multivariate Permutation Tests: With Applications in Biostatistics*, retitled *Permutation Tests for Complex Data*, by Fortunato Pesarin and Luigi Salmaso in 2010 [1122]; and *Mathematical Statistics with Resampling and R* by Laura Chihara and Tim Hesterberg in 2011 [253].

The journal articles on permutation methods published between 2001 and 2010 are too numerous to be summarized in any detail. A search of The Web of Science® for "permutation" lists 9,259 journal articles and 73,960 citations for this period, with steady increases for each year. For example, in 2000 there were 1,619 citations, in 2005 there were 5,862 citations, and in 2010 there were 15,612 citations. The journal articles may be conveniently divided into two areas: fields of research and research methods.

A cursory examination of the fields of research in which articles using permutation methods were published includes atmospheric science, bioinformatics, biology, chemistry, clinical trials, cognition, computer science, conservation, ecology, environmental research, epidemiology, forestry, genetics, geology, history, industrial engineering, medicine, molecular biology, operations research, physiology, public health, statistics, and veterinary medicine.

The research methods for which permutation tests were published in this period included, but were not limited to, multiple regression, analysis of variance, canonical correlation, quantile regression, the Wilcoxon and Mann–Whitney two-sample rank-sum tests, the Jonckheere–Terpstra test, trend analysis, matched pairs, analysis of multivariate data, partitions, Cohen's kappa measure of agreement, categorical variation, ordered and unordered contingency tables, qualitative variation, survival analysis, Cronbach's alpha, tetrachoric correlation, ridit analysis, discriminant analysis, and robustness.

Finally, it should be mentioned that the computer language, R, first released in 1995, became immensely popular among statisticians in this period. R is an open-source programming language designed especially for statistical computing and graphics, and was first developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. R is more than a programming language and is best described as an interactive environment for doing statistics. Thus, because R provides a wide variety of statistical procedures, including linear and nonlinear modeling, classical statistical tests, time-series analysis, and others, R quickly became the language of choice after 2000 for many researchers. Moreover, R is highly extensible and is easy to program for both exact and resampling permutation methods.

6.4 A Summary of Contributions by Publication Year

In this section a brief description of representative articles published after 2000 is provided, organized by year of publication. While the number of articles published in this period is too large to summarize completely, a sample of articles will convey the flavor of the times. The articles on permutation methods published between 2001 and 2010 are heavily concentrated in four general areas: linear correlation and regression, analysis of variance, measures of agreement and concordance, and contingency table analysis.

Linear Correlation and Regression. On this topic are articles by Huh and Jhun [669] and Anderson and Robinson [21] in 2001; Mielke and Berry [963] and Sakaori [1215] in 2002; O'Gorman [1050] in 2005; Yamada and Sugiyama [1471] and Cade and Richards [234] in 2006; Long, Berry, and Mielke [840] in 2007; Önder [1062] in 2008; and Long, Berry, and Mielke [841] in 2009.

Analysis of Variance. In the general area of analysis of variance are articles by Weinberg and Lagakos [1425] in 2001; Pesarin and Salmaso [1121] in 2002; Jin and Robinson [687], Graves, Reese, and Fitzgerald [546], and Anderson and ter Braak [19] in 2003; Ernst [413] in 2004; Raab and Butcher [1148] and Good [531] in 2005; Jung, Jhun, and Song [703], Corain and Salmaso [277], Wheldon, Anderson, and Johnson [1440], Kaiser [705], and Önder [1061] in 2007; Good and Xie [533] and Fraker and Peacor [476] in 2008; Finch and Davenport [433] and Zhang [1491] in 2009; Reiss, Stevens, Shehzad, Petkova, and Milham [1163] and Mewhort, Johns, and Kelly [928] in 2010; and Berry, Johnston, and Mielke [117] in 2011.

Agreement and Concordance. On the topic of agreement and concordance are articles by Berry and Mielke [157] in 2001; Legendre [810] and Berry, Johnston, and Mielke [112] in 2005; Norman and Scott [1043] in 2007; Brusco, Stahl, and Steinley [226] and Mielke, Berry, and Johnston [976] in 2008.

Contingency Table Analysis. In the general area of contingency table analysis are articles by Agresti [3] in 2001; Cryan and Dyer [299] in 2003; Borkowf [183] in 2004; Berry, Johnston, and Mielke [113] in 2006; Campbell [239] in 2007; Long, Berry, and Mielke [841], Hitchcock [633], and Mielke, Long, Berry, and Johnston [986] in 2009.

Permutation Methods in 2001

A number of notable articles on permutation methods were published in 2001. Agresti published an influential overview article in *Statistics in Medicine* on exact inferences for categorical data with special attention to the interval estimation of a proportion and the odds-ratio statistic [3]. Huh and Jhun developed an alternative random-permutation testing method for multiple linear regression [669]. They claimed in this article that the new method was an improvement over the methods previously proposed by Freedman and Lane in 1983 [478] and by Kennedy in 1995 [748]. Anderson and Robinson observed that there was general agreement concerning an appropriate method for exact tests of hypotheses in simple linear regression [21]. However, this was not the case, they noted, for partial tests in multiple regression, citing papers on the topic by Brown and Maritz [224], Freedman and Lane [478], Collins [269], Gail, Tan, and Piantadosi [491], Kennedy [748], Manly [876], Oja [1052], ter Braak [1346], and Welch [1432]. Anderson and Robinson compared the distributions of test statistics under various permutation methods proposed via simulation [21]. Two articles in 2001 on permutation tests for multiple linear regression by Huh and Jhun [669] and by Anderson and Robinson [21] were harbingers of the many articles on this topic in the next 10 years. Finally in 2001, Weinberg and Lagakos compared rank and permutation tests based on summary statistics computed from repeated-measures (blocked) data [1425]. They used recent theoretical results for the non-null behavior of rank and permutation tests to examine the asymptotic relative efficiencies of several popular summary statistics.

Permutation Methods in 2002

In 2002 extensions of multiple regression permutation analyses to applications involving multivariate dependent values were considered by Mielke and Berry [963]; see also [964]. The extensions were prompted by a multivariate multiple regression algorithm developed by Kaufman, Taylor, Mielke, and Berry in 2002 [711]. Sakaori investigated permutation tests for equality of correlation coefficients between two independent populations [1215]. He demonstrated how to apply a permutation test to the problem and discussed its asymptotic suitability. Finally in 2002, Pesarin and Salmaso published an article exploring exact permutation testing of effects in unreplicated two-level multifactorial designs [1121]. The approach provided by Pesarin and Salmaso preserved the exchangeability of error components for testing up to k effects in 2^k designs (q.v. page 4). They further discussed the advantages and limitations of exact permutation procedures and executed a simulation study utilizing the Iris data of Fisher based on a paired-permutation strategy.[3]

Permutation Methods in 2003

In 2003 Cryan and Dyer developed a polynomial-time algorithm to approximate the number of possible arrangements of cell frequencies in a contingency table when the number of rows is constant [299]. Also in 2003, Jin and Robinson published an article on robust permutation tests for one sample [687]. Specifically, they considered robust permutation tests based on an estimating equation comparing test statistics based on the score function with those based on the M estimator. Graves, Reese, and Fitzgerald proposed a new class of models for permutations based on a Bayesian hierarchical framework, which permited hierarchical specification and fully hierarchical estimation of interaction terms [546]. Janssen and Pauls published a lengthy article in *The Annals of Statistics* titled "How do bootstrap and permutation tests work?" [680]. This was an ambitious paper of 40 pages that considered a comprehensive and unified approach for the conditional and unconditional analysis of linear resampling permutation statistics. Finally in 2003, Anderson and ter Braak published an article that provided guidelines for constructing an exact permutation strategy, where possible, for any individual term in any analysis of variance design (e.g., fixed, mixed, random, or nested) [19]. In addition, Anderson and ter Braak provided results of Monte Carlo simulations to compare the level of accuracy and power of different permutation strategies in two-way analysis of variance designs, including mixed models, nested hierarchies, and tests of interaction [19].

Permutation Methods in 2004

In 2004 Ernst published a review article in *Statistical Science* on permutation methods as a basis for exact inference [413]. This comprehensive overview article provided an extensive introduction to permutation methods and included exact inference procedures, hypothesis tests, confidence intervals, both the permutation and population models (q.v. page 3), two-group permutation tests, one-way analysis of variance, and multiple comparisons of means tests. Borkowf presented an algorithm for generating two-way contingency tables with fixed marginal frequency totals and

[3] The Fisher Iris data is a multivariate data set analyzed by Fisher in 1936 to illustrate discriminate analysis [454]. The data were collected and originally published by Edgar Anderson in 1935 and 1936 [17, 18].

arbitrary mean proportions [183]. As Borkowf noted, such tables have exactly the multivariate extended hypergeometric (MXH) distribution and have many important applications. Finally in 2004, Good wrote a letter to the editor of *Statistics in Medicine* in an attempt to clarify some confusion about the nature of permutation tests, as suggested by the titles and contents of some recent articles published in *Statistics in Medicine* [529].[4] In this letter Good promoted permutation tests over rank tests, observing that the great value of permutation tests in medical research lies in their power, robustness, and ability to provide exact probability values.

Permutation Methods in 2005

In this period there was sustained interest in developing and enhancing measures of agreement and concordance;permutation versions of various agreement measures accompanied this interest. Thus, many publications appeared dealing with measures of agreement and concordance.In 2005 Legendre published an interesting article on Kendall's measure of concordance[810] and Berry, Johnston, and Mielke expanded on their previously-published work on Cohen's kappa measure of agreement, both unweighted and weighted, expanding kappa to measure agreement among multiple raters [112,114,115]. Raab and Butcher discussed the choice of randomization tests for inference from cluster-randomized trials that have been designed to ensure a balanced allocation of clusters to treatments [1148]. Two cluster-randomized trials with balanced designs were used to illustrate the possible choices in selecting a randomization test, and methods for obtaining confidence intervals for treatment effects were illustrated. One cluster-randomized trial conducted in the Lothian and Tayside regions of southeast Scotland evaluated whether a specially developed teacher-led sex education program delivered in Scottish schools had any effect on unsafe sexual behaviors, unwanted sexual outcomes, and the quality of sexual relationships. The second cluster-randomized trial conducted in the Northern and Yorkshire regions of England investigated whether a dietician-led training program in obesity management for primary care teams resulted in changes in weights of patients.

O'Gorman evaluated the performance of randomization tests that use permutations of independent variables for a subset of regression coefficients in a linear model [1050]. O'Gorman showed that permuting the independent variables maintained the level of significance and possessed power that approximated the power of randomization tests based on permutation of residuals from a reduced regression model. In 2005 Good wrote another letter to the editor of *Statistics in Medicine* regarding the "efficiency comparisons of rank and permutation

[4] Although there are no references in the letter by Good, it is readily apparent that Good was basing his criticisms on an article by Weinberg and Lagakos titled "Efficiency comparisons of rank and permutation tests based on summary statistics computed from repeated-measures data," which was published in *Statistics in Medicine* in 2001 [1425].

tests" [530].[5] In this second letter, like the first, Good made a case for the use of rank tests in two situations: when outliers are a concern or when combining results with different precision. Specifically, Good observed that as a test based on ranks, the Wilcoxon two-sample rank-sum test does not use all the available information, in contrast to the Fisher–Pitman permutation test. However, Good argued the Wilcoxon two-sample rank-sum test could be more powerful for asymmetrical and heavy-tailed distributions. In support of this position, Good cited earlier works by Keller-McNulty and Higgins in 1987 [714] and van den Brink and van den Brink in 1989 [1389]. Good concluded that the Wilcoxon two-sample rank-sum test can be much more powerful than the Fisher–Pitman permutation test under non-normality. In fact, he contended that the Fisher–Pitman permutation test achieved only a modest power advantage over the Wilcoxon two-sample rank-sum test for the normal distribution. Finally, Good argued that the Wilcoxon two-sample rank-sum test is a good choice when the underlying distribution function is a priori unknown [529, 530].[6]

Permutation Methods in 2006

In 2006 Yamada and Sugiyama developed a permutation test statistic for canonical correlation analysis, establishing that the permutation test possessed more power than the conventional asymptotic test [1471]. Also in 2006, Cade and Richards published a permutation test for quantile regression [234]. They observed that estimating the quantiles of a response variable conditioned on a set of covariates in a linear model has many applications in the biological and ecological sciences, as quantile regression models allow the entire conditional distribution of a response variable y to be related to some covariates $\mathbf{X}$, providing a richer description of functional changes than is possible by focusing on just the mean or other central statistics [234, p. 106].

Permutation Methods in 2007

In 2007 Campbell added to the protracted examination and re-examination of the Pearson chi-squared and Fisher–Irwin exact tests for 2×2 contingency tables [239]. Also in 2007, a long-standing problem was solved when an efficient resampling algorithm was developed for multi-way contingency tables, thereby enabling multidimensional permutation analyses of various problems, including

[5] The first letter was published in *Statistics in Medicine* in 2004 [529].

[6] It should be noted that Good took a position on the use of rank tests in stark contrast to other researchers of the time, many of whom were abandoning rank tests in favor of permutation tests using the original raw score measurements instead of converting raw scores to rank-order statistics (q.v. page 402).

Cohen's kappa measure of agreement with multiple raters [975, 976]. Jung, Jhun, and Song developed an exact permutation method for testing both interaction and main effects in two-way analysis of variance models that they concluded was an improvement over previous methods such as presented by Still and White in 1981 [1324] and ter Braak in 1982 [703]. Gill showed that the exact probability values of permutation and bootstrap hypothesis tests of differences among groups could be written as an infinite series whose terms could be rapidly computed [513]. This same technique was later to be used by Mewhort, Johns, and Kelly in an analysis of factorial designs in 2010 [928]. Corain and Salmaso published a critical review and comparative study regarding conditional permutation tests for the two-way analysis of variance [277], and Norman and Scott demonstrated the adverse effects of serially-observed data sequences containing transient events on the calculation of Cohen's index of inter-rater agreement [1043]. They developed a Monte Carlo permutation procedure to produce an empirical distribution of Cohen's kappa in the presence of serial dependence.

Wheldon, Anderson, and Johnson proposed a new procedure for the analysis of the large, multi-dimensional data arrays produced by electroencephalographic (EEG) measurements of human brain function [1440]. They proposed a three-step approach whereby they (1) summed univariate statistics across variables, (2) used permutation tests for treatment effects at each point in time, and (3) adjusted for multiple comparisons using permutation distributions to control for family-wise error. Kaiser derived Monte Carlo simulations for the Fisher–Pitman permutation tests for paired replicates and independent samples [705], developing algorithms and providing Stata implementations for both tests. Finally in 2007, Önder used permutation tests to reduce type I and type II errors in small ruminant research [1061].[7] He concluded on the basis of several analyses that "permutation tests should be preferred to t and F tests to avoid type I and II errors" [1061, p. 72].

Permutation Methods in 2008

In 2008 Brusco, Stahl, and Steinley developed an implicit enumeration algorithm for an exact test of weighted kappa [226]. Also in 2008, H.A. David, writing in *The American Statistician*, provided a concise history of the early beginnings of permutation methods [326]. Good and Xie analyzed a balanced crossover design with permutation rather than parametric methods in order to obtain exact distribution-free significance levels that were independent of the underlying distribution, thus controlling type I error and increasing power [533]. They then showed how the permutation method could be extended to any number of treatment sequences and treatments in a balanced crossover design. Önder published a comparative study of permutation tests with Euclidean and Bray–Curtis distances for common agricultural distributions in regression [1062]. He examined normal, Poisson,

[7] The small ruminants studied by Önder were purebred Ile de France sheep and crossbred Chios and Awassi sheep.

chi-squared, and Cauchy distributions, concluding that permutation of the raw data with Euclidean distance was to be recommended when the sample size was less than 15 and permutation of the residuals under the full model with Euclidean distance was preferred with samples larger than 15 for all distributions, except the normal distribution. Önder found that Bray–Curtis distances were simply not suitable for the four distributions examined.[8] Jiang and Kalbfleisch published a paper on permutation methods in relative-risk regression models, a new application for permutation methods [686]. In this paper they developed a weighted permutation method to construct confidence intervals for regression parameters in relative-risk regression models. A simulation study established that the weighted permutation method typically improved accuracy over conventional asymptotic confidence intervals.

Finally in 2008, Fraker and Peacor compared permutation tests and the conventional analysis of variance in testing for biological interactions [476]. Noting that interaction terms from statistical tests are often used to make inferences about biological processes, they argued that it is critical that the statistical method that is used tests a model that corresponds to a realistic biological null hypothesis. As examples of biological interactions, they offered, first, when a predator interacts with a consumer to affect resource density via predator-induced changes in consumer behavior (behaviorally mediated trophic cascades) and, second, when the probability of certain species of plants becoming established in new areas depends on the presence of other species (facilitative interactions) [476]. Fraker and Peacor provided two simulated experiments of species interactions. With some caveats, they concluded that permutation tests provide an advantage over the conventional analysis of variance in their ability to test a wider range of models and should be used to make inferences concerning biological interactions [476].

Permutation Methods in 2009

In 2009 Long, Berry, and Mielke developed a permutation alternative to tetrachoric correlation [841], and Mielke, Long, Berry, and Johnston extended the classical two-treatment ridit analysis first introduced by Bross in 1958 to $g \geq 2$ treatments utilizing a resampling-approximation permutation procedure to obtain approximate upper-tail probability values [986]. Also in 2009, Hitchcock published a comprehensive review of Frank Yates and his work on contingency tables, with special attention to the controversial 1934 correction for continuity [633]. Knijnenburg, Wessels, Reinders, and Shmulevich developed a method for computing approximate permutation probability values by using a generalized Pareto distribution [761].

LaFleur and Greevy wrote a methodological article in *Journal of Clinical Child & Adolescent Psychology* that introduced exact and resampling-approximation permutation methods to that discipline [789]. They used an application-based approach

[8] Technically, Bray–Curtis is a dissimilarity measure, not a distance measure, as it does not satisfy the triangle inequality (q.v. page 255).

to provide a tutorial on permutation testing, presenting some historical perspectives, describing how permutation tests are formulated, providing examples of research situations under which permutation methods are useful, and demonstrating the utility of permutation methods to clinical and adolescent psychologists [789].

Finch and Davenport, writing in *Methodology*, explored the performance of Monte Carlo permutation and approximate tests for multivariate means comparisons with small sample sizes when parametric assumptions are violated [433]. A simulation study compared the performance of four standard multivariate analysis of variance (MANOVA) test statistics (Wilks' likelihood-ratio test, the Bartlett–Nanda–Pillai Trace, the Lawley–Hotelling Trace, and Roy's Maximum Root) with their Monte Carlo permutation-based counterparts under a variety of conditions with small samples, including conditions when the assumptions underlying MANOVA were met and when the assumptions were not met [965, pp. 53–57]. They concluded that under conditions similar to those presented in the paper, the four approximate F tests and Monte Carlo permutation versions of the tests used in MANOVA performed similarly across a variety of conditions. Finally, they noted that Roy's Maximum Root provided a dramatic improvement over the F approximation obtained from a standard MANOVA analysis and suggested that researchers should consider using it in practice [433, p. 68]. Huang, Jin, and Robinson considered robust permutation tests for a location shift in the two-sample case based on estimating equations and compared the results with those of test statistics based on a score function and an M estimator [662].

Arboretti Gianchristofaro, Bonnini, and Pesarin developed a permutation approach for testing heterogeneity in two-sample categorical variables, wherein the problem is one of establishing whether the distribution of a categorical variable is more concentrated—less heterogeneous—in one or the other of two populations [29]. Yu, Kepner, and Iyer, writing in *Biometrical Journal*, proposed an exact permutation method with respect to testing cytostatic cancer treatment using correlated bivariate binomial random variables to simultaneously assess two primary outcomes: for advanced Hepatocellular carcinoma measures of therapy efficiency in terms of disease control rate and no progression rate at six months [1479]. Finally in 2009, Zhang proposed a new type of permutation procedure for testing the difference between two population means: split-sample permutation t tests that do not require the exchangeability assumption (q.v. page 4), are asymptotically exact, and can easily be extended to testing hypotheses about a single population [1491].

Permutation Methods in 2010

In 2010 Mewhort, Johns, and Kelly [928] noting that permutation tests are seldom applied to factorial designs because of the computational load they impose, proposed two methods to limit the computational burden. First, they showed that using orthogonal contrasts greatly reduced the number of computations required and second, that when combined with a new algorithm by Gill [513], a permutation test for factorial designs was both practical and efficient [928]. Also in 2010, Reiss, Stevens, Shehzad, Petkova, and Milham compared pseudo-F tests and multi-response permutation procedures (MRPP) in assessing multivariate observations [938, 965, 1163]. They further showed under what conditions the two procedures were identical.

6.5 Agresti and Exact Inference for Categorical Data

In 2001 Alan Agresti published a lengthy article in *Statistics in Medicine* on recent advances and continuing controversies associated with exact inference for categorical data [3]. This article was, in part, an overview article, but also one that examined and summarized some of the criticisms of exact methods, e.g., the conservative nature of exact methods because of the inherent discreteness of the permutation distribution.

Agresti began his article with a formal introduction to the exact conditional approach for categorical data arranged in a $r \times c$ contingency table. Next, he discussed the exact unconditional approach for categorical data, beginning with the comparison of binomial parameters for two independent samples.[9] Agresti noted in this section of his paper that statisticians have been critical of both exact conditional and exact unconditional approaches, a theme that appeared and reappeared in the past and is still not settled. Critics of the exact conditional approach disagree with analyzing sample spaces consisting only of contingency tables with exactly the same marginal frequency totals as the observed contingency table. Proponents of the exact conditional approach, however, respond that it is unnatural to consider samples that differ from the marginal distributions of the observed contingency table. Still other statisticians have argued that the exact unconditional approach is artificial as it averages what happened in the observed sample with hypothetical response distributions [3, pp. 2712–2713]. In this regard, Agresti cited previous articles on the topic by Suissa and Shuster [1333, 1334], Little [836], Routledge [1197], Greenland [551], Upton [1386], Martín Andrés [899], Reid [1161], Howard [655], Cormack and Mantel [281], and Yates [1476].[10] Agresti concluded that much of the disagreement is due to the quite different results that the two approaches can yield when the distribution is highly discrete, e.g., when the sample size is small.

Agresti devoted two sections of the paper to complications from discreteness, illustrating the problem with numerous examples involving samples with small sample sizes. He explained that in the real world it is rarely possible to achieve an arbitrary critical value such as $\alpha = 0.05$ with randomization, noting that some argue that fixing an unachievable α level is artificial and that one should merely report

[9] Agresti's "exact conditional approach" corresponds to marginal frequency totals that are fixed, while his "exact unconditional approach" corresponds to marginal frequency totals that are not fixed.

[10] See also an informative 2011 article by Yung-Pin Chen comparing the chi-squared and Fisher's exact probability tests in *The American Statistician* [251].

the probability value. However, he countered, it is more important in constructing confidence intervals, as one knows only that the actual confidence level is at least as high as, say, 95 %; one does not know the actual level, since one does not know the true parameter value [3, p. 2715].

Finally, Agresti offered a compromise: use adjustments of exact methods based on the mid-P value, as advocated by Lancaster [794].[11] The mid-P procedure uses one-half the probability of the observed contingency table, plus the probability values of those contingency tables that are less than that of the observed contingency table. Agresti argued that inference based on the mid-P method appears to be a sensible accommodation between the conservativeness of exact methods and the uncertainty of large-sample methods. The article concluded with an extensive listing of 98 references that is invaluable for researchers interested in this topic.

6.6 The Unweighted Kappa Measure of Agreement

In 2001 Berry and Mielke compared two popular measures of agreement: Cohen's κ_c [263] and Brennan and Prediger's κ_n [210], using exact and resampling-approximation permutation methods [157].[12] Consider an $r \times r$ agreement matrix. Then, Cohen's test statistic is given by

$$\kappa_c = \frac{\displaystyle\sum_{i=1}^{r} p_{ii} - \sum_{i=1}^{r} p_{i.}p_{.i}}{1 - \displaystyle\sum_{i=1}^{r} p_{i.}p_{.i}},$$

where i denotes the ith of r categories, p_{ii} indicates the observed proportion of agreements in row i and column i, and $p_{i.}$ ($p_{.i}$) is the proportion of objects assigned to category i by Judge 1 (2). In contrast, Brennan and Prediger's test statistic is given by

$$\kappa_n = \frac{\displaystyle\sum_{i=1}^{r} p_{ii} - \frac{1}{r}}{1 - \dfrac{1}{r}}.$$

In 1988 Zwick [1498] observed that Brennan and Prediger's proposed measure was not new and had previously been termed the S coefficient by Bennett, Alpert, and Goldstein in 1954 [95], the C coefficient by Janson and Vegelius in 1979 [679], and

[11] See also a comprehensive review of the use of the mid-P procedure by Berry and Armitage in *The Statistician*, published in 1995 [107].

[12] In this section, Cohen's kappa is indicated by κ_c to distinguish it from the κ_n of Brennan and Prediger.

Fig. 6.1 Data on $n = 38$ observations classified into $r = 3$ categories by two independent raters

	A	B	C	Total
A	24	0	0	24
B	3	1	3	7
C	7	0	0	7
Total	34	1	3	38

in the case of $r = 2$, the G index by Holley and Guilford in 1964 [642], as well as the random error (RE) coefficient by Maxwell in 1977 [907].

Obviously, the two measures, κ_c and κ_n, provide different interpretations of agreement between two judges. Cohen's κ_c measure corrects the raw proportion of agreement, $\sum_{i=1}^{r} p_{ii}$, for chance agreement by $\sum_{i=1}^{r} p_{i.}p_{.i}$, whereby the marginal distributions are taken into consideration. On the other hand, Brennan and Prediger's κ_n measure corrects the raw proportion of agreement by $1/r$, which is the average value of $\sum_{i=1}^{r} p_{ii}$ for all possible $r \times r$ agreement matrices given the sample size, n, and ignoring the marginal distributions.

As noted by Berry and Mielke, it is relatively straightforward to compute an exact probability value for κ_n [157]. Because $1/r$ is a known constant, the distribution of κ_n is equivalent to the distribution of $\sum_{i=1}^{r} p_{ii}$. If

$$y = n \sum_{i=1}^{r} p_{ii} ,$$

then y follows a binomial distribution and the probability (P) value for κ_n is given by

$$P(\kappa_n) = \sum_{i=y}^{n} \binom{n}{i} \left(\frac{1}{r} \right)^{i} \left(1 - \frac{1}{r} \right)^{n-i} .$$

It is considerably more difficult to obtain a probability value for κ_c. Berry and Mielke took a unique approach, converting the raw data to a randomized block design with n observers, two blocks, and the r categories represented by an $r \times 1$ binary vector, where the ith element, corresponding to the ith of r categories was set to 1 and the remaining $r - 1$ elements were set to zero. They then relied on a resampling-approximation algorithm whereby a sample of L random permutations was extracted from the $(n!)^2$ possible permutations and the desired probability value was the proportion of the L values of κ_c equal to or greater then the observed value of κ_c [157].

To illustrate the block-design approach, consider an example of $n = 38$ observations, each classified into one of $r = 3$ mutually exclusive categories (A–C) by two independent judges, as depicted in Fig. 6.1. The results for the data listed in Fig. 6.1 are quite different: $\kappa_n = +0.4868$ with an upper-tail binomial probability value of 0.4330×10^{-4}, and $\kappa_c = +0.1767$ with an upper-tail resampling-approximation probability value of 0.7218×10^{-1}, based on $L =$ 1,000,000 random permutations.

6.7 Mielke et al. and Combining Probability Values

Exact permutation tests are based on all possible arrangements of observed data sets; consequently, exact permutation tests yield probability values obtained from discrete probability distributions. In 2004 Mielke, Johnston, and Berry introduced an exact nondirectional method to combine independent probability values that obey discrete probability distributions [985]. A nondirectional method developed by R.A. Fisher in 1925 is known to possess excellent asymptotic properties for combining independent probability values from continuous uniform distributions [448]; see also two articles on this topic by Littell and Folks in *Journal of the American Statistical Association* in 1971 and 1973 [834, 835]. The purpose of the paper by Mielke et al. was to introduce a discrete analog of Fisher's classical method to combine independent probability values from permutation tests, a necessary component for conducting meta-analyses of research based on permutation methods.

Fisher's 1925 method for combining k independent probability values $(P_1, \ldots, P_k)$ from continuous probability distributions is based on the statistic

$$T = -2 \ln \left(\prod_{i=1}^{k} P_i \right) = -2 \sum_{i=1}^{k} \ln(P_i),$$

which is distributed as chi-squared with $2k$ degrees of freedom, under the null hypothesis that $P_1, \ldots, P_k$ are independent uniform random variables between 0 and 1. If T_{o} denotes the observed value of T, then Fisher's combined probability value is $P(T \geq T_{\text{o}}|H_0)$. Consequently, Fisher's classical method is not appropriate for independent probability values obeying discrete probability distributions where only a limited number of different events are possible [985, p. 450]; see also articles on this topic by H.O. Lancaster in 1949 [792] and E.S. Pearson in 1950 [1096].

The method proposed by Mielke et al. is conceptually applicable to obtaining exact combined probability values for a multitude of independent tests, including the Fisher exact probability test, exact chi-squared and exact likelihood-ratio tests, the Fisher–Pitman permutation test, and rank tests such as the Ansari–Bradley test [26], Mood's median test [1001], the Taha test [1339], and the Wilcoxon–Mann–Whitney two-sample rank-sum test [880, 1453]. The method is applied here to the Fisher exact probability test.

Following the notation of Mielke et al., let $p_{ij} > 0$ denote the point-probability value for the ith of k specified discrete probability distributions to be combined and the jth of m_i events associated with the ith discrete probability distribution; thus, $i = 1, \ldots, k$ and $j = 1, \ldots, m_i$. Also let $p_{i\,\text{o}}$ denote the observed probability value of p_{ij}. Under H_0, the exact combined probability value of the k discrete probability distributions for Fisher's exact probability test is given by

$$\sum_{j_1=1}^{m_1} \cdots \sum_{j_k=1}^{m_k} \alpha_{j_1}, \ldots, _{j_k} \prod_{i=1}^{k} p_{ij_i},$$

where

$$\alpha_{j_1,\ldots,j_k} = \begin{cases} 1 & \text{if } \prod_{i=1}^{k} p_{ij_i} \leq \prod_{i=1}^{k} p_{i\mathrm{o}}, \\ 0 & \text{otherwise.} \end{cases}$$

Example analyses of a variety of sparse 3 × 4 contingency tables demonstrated that Fisher's continuous method was not appropriate for discrete probability distributions from sparse data tables. Mielke et al. further established that the inclusion of even a single discrete probability distribution could have a substantial negative effect on Fisher's continuous method to combine probability values [985, p. 456]. In 2005 Mielke, Berry, and Johnston extended the method for combining independent probability values from discrete distributions to combining probability values associated with the permutation version of the matched-pairs t test [972].

6.8 Legendre and Kendall's Coefficient of Concordance

In 2005 Legendre published an article on the Kendall coefficient of concordance (W), with application to species association [810]. Specifically, Legendre utilized a permutation version of Kendall's coefficient of concordance to identify groups of significantly associated species of oribatid (beetle) mites in the peat blanket surrounding a bog lake.[13] He noted that $p(n-1)W$ is asymptotically distributed as chi-squared with $n-1$ degrees of freedom, where p and n denote the number of species and number of sites, respectively.[14] However, the distribution of species across sites was highly skewed, exacerbated by a high frequency of zero values. Therefore, while the classical chi-squared test would be overly conservative, a permutation test would have the correct type I error [810, p. 226].

Legendre permuted the sites 10,000 times and calculated the rate of rejection of the null hypothesis (the p species produced independent rankings of the sites), together with 95 % confidence intervals. Legendre assessed the contribution of individual species by a modified permutation test, noting that in a permutation framework a post hoc test of the contribution of each species to the overall W concordance statistic is possible, which is not the case in the classical testing framework [810, p. 226].

Prior to the concordance analysis, the abundance data on the oribatid mites were transformed using a Hellinger transformation. The Hellinger transformation consists of two steps: (1) express each abundance value as a proportion with respect to the total sum of animals collected at a site, and (2) take the square root of that

[13] The oribatid mite is considered to be the world's strongest animal, able to support 1,180 times its weight. By contrast, the strongest human can support approximately three times its weight.

[14] Here, the number of species (p) is the number of judges or raters in the usual implementation of Kendall's coefficient of concordance.

proportion [810, p. 228]. As Legendre explained, such a transformation ensures that the Euclidean distance computed among sites for the transformed data is equal to the Hellinger distance for the untransformed data.[15]

Based on extensive simulation, Legendre observed that when the null hypothesis was true, permutation testing lead to correct type I error in tests of significance of the Kendall coefficient of concordance. Moreover, in the classical chi-squared test, type I error was too low when the number of species (judges) was less than 20, leading to tests that were overly conservative with reduced power. Finally, Legendre concluded that because in most real-life applications the number of species is small "permutation tests should be routinely used to test Kendall's W statistic" [810, p. 243].

6.9 The Weighted Kappa Measure of Agreement

In 2005 Berry, Johnston, and Mielke turned attention to Cohen's weighted kappa statistic, utilizing exact and resampling-approximation procedures [112]. Weighted kappa differs from unweighted kappa in that weights are assigned to cells and progress outward from the agreement diagonal in an $r \times r$ cross-classification table [264]. Thus, disagreements among judges can be weighted or scaled, providing greater weights for more serious disagreements. Two weighting schemes are widely used in weighted kappa: quadratic weighting where the weights are given by $w_{ij} = (i - j)^2$ for $i, j = 1, \ldots, r$, and linear weighting where $w_{ij} = |i - j|$ for $i, j = 1, \ldots, r$. For unweighted kappa, the weights are given by

$$w_{ij} = \begin{cases} 0 & \text{if } i = j, \\ 1 & \text{otherwise.} \end{cases}$$

For detailed discussions regarding choices of weights, see articles by Maclure and Willett in 1987 [864]; Graham and Jackson in 1993 [544]; Banerjee, Capozzoli, McSweeney, and Sinha in 1999 [60]; Kundel and Polansky in 2003 [781]; and, especially, Schuster and Smith in 2005 [1240].

J.E. Johnston

Janis E. Johnston received B.S. degrees in mathematics and natural science from the University of Wyoming in 1994, her M.A. in sociology from the University of Wyoming in 1999, and her Ph.D. in sociology from Colorado State University in 2006. From 2007 to 2009 she was a Science & Technology Policy Fellow with the American Association for the Advancement of Science (AAAS), where she worked with the Environmental Protection Agency

(continued)

[15] For more on the Hellinger and other transformations of species data, see a 2001 article by Legendre and Gallagher [811].

(EPA), National Homeland Security Research Center (NHSRC). She has been a Social Science Policy Analyst with the United States government in Washington, DC since 2010.

Assume that two independent judges have assigned each of n objects to one of an exhaustive set of r mutually exclusive categories. The ratings of the two judges are cross-classified into an $r \times r$ contingency (agreement) table where n_{ij} denotes the frequencies with which two judges assigned objects to the ith and jth categories, respectively, for $i, j = 1, \ldots, r$. Let $n_{i.}$ denote the total number of objects assigned to the ith category, $n_{.j}$ denote the total number of objects assigned to the jth category by the first and second judge, and n denote the total number of objects, respectively. Then, Cohen's weighted kappa test statistic is given by

$$\hat{\kappa} = 1 - \frac{S}{S_e},$$

where

$$S = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} p_{ij}$$

is the proportion of weighted disagreement between the two judges,

$$S_e = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} p_{i.} p_{.j}$$

is the proportion of disagreements expected by chance under the null hypothesis, $p_{ij} = n_{ij}/n$, $p_{i.} = n_{i.}/n$, $p_{.j} = n_{.j}/n$ are the observed cell and marginal proportions, w_{ij} denotes the disagreement weights, and the null hypothesis specifies $p_{ij} = p_{i.} p_{.j}$ for $i, j = 1, \ldots, r$. Thus, S_e is the expected value of S under the null hypothesis [112].

In the context of an $r \times r$ contingency table with n objects cross-classified by the ratings of two independent judges, an exact permutation test requires enumeration of all possible arrangements of objects to the r^2 cells, while preserving the marginal frequency totals. For each arrangement of cell frequencies, the weighted kappa test statistic, $\hat{\kappa}$, and the exact hypergeometric probability, $P(n_{ij}|n_{i.}n_{.j})$, are calculated, where

$$P(n_{ij}|n_{i.}n_{.j}) = \frac{\left(\prod_{i=1}^{r} n_{i.}!\right)\left(\prod_{j=1}^{r} n_{.j}!\right)}{n! \prod_{i=1}^{r} \prod_{j=1}^{r} n_{ij}!}.$$

If $\hat{\kappa}_{\text{o}}$ denotes the value of the observed weighted kappa test statistic, the exact one-sided upper- and lower-tail probability values of $\hat{\kappa}_{\text{o}}$ are the sums of the $P(n_{ij}|n_{i.}n_{.j})$ values associated with those $\hat{\kappa}$ values equal to or greater than $\hat{\kappa}_{\text{o}}$ and equal to or less than $\hat{\kappa}_{\text{o}}$, respectively. Small upper-tail probability values usually imply agreement, whereas small lower-tail probability values usually imply disagreement [112, p. 246]. Note that the hypergeometric probability values, $P(n_{ij}|n_{i.}n_{.j})$, are accumulated with respect to the ordered weighted kappa values [511, 650, 1150].

When the number of possible arrangements of cell frequencies is very large, exact permutation tests are impractical and permutation tests based on resampling become necessary. Berry, Johnston, and Mielke utilized a resampling algorithm to generate random arrangements of cell frequencies from a two-way contingency table, given fixed marginal frequency totals, as described by Patefield [1089] (q.v. page 281). The resampling one-sided upper- and lower-tail probability values of $\hat{\kappa}_{\text{o}}$ were simply the proportions of the resampled $\hat{\kappa}$ values equal to or greater than $\hat{\kappa}_{\text{o}}$ and equal to or less than $\hat{\kappa}_{\text{o}}$, respectively.[16] Finally, Berry et al. noted that when testing the null hypothesis that the population value of weighted kappa is zero, then

$$Z = \frac{\hat{\kappa}}{\sigma_{\hat{\kappa}}}$$

is approximately distributed as $N(0,1)$ and the asymptotic one-sided upper- and lower-tail probability values are given by $P(Z \geq \hat{\kappa}_{\text{o}}/\sigma_{\hat{\kappa}})$ and $P(Z \leq \hat{\kappa}_{\text{o}}/\sigma_{\hat{\kappa}})$, respectively, where $\sigma_{\hat{\kappa}}$ was calculated using an exact variance algorithm first described by Everitt in 1968 [415] and reformulated into a form favorable to computation by Mielke, Berry, and Johnston in 2005 [973] (q.v. page 394).

6.10 Berry et al. and Measures of Ordinal Association

In 2006 Berry, Johnston, and Mielke introduced efficient permutation procedures for exact and resampling one-sided probability values for six popular measures of association between two ordered variables: Kendall's τ_a and τ_b [736], Stuart's τ_c [1326], Goodman and Kruskal's γ, [534], and Somers' d_{yx} and d_{xy} [1294].

Following the notation of Berry et al., consider two ordinal variables, X and Y, cross-classified into an $r \times c$ contingency table, where r and c denote the number of rows and columns, respectively. Let n_{ij} denote the number of objects in the ijth cell, $i = 1, \ldots, r$ and $j = 1, \ldots, c$, and let n denote the total number of objects in the $r \times c$ table, i.e.,

[16] Note that, in contrast to an exact permutation test, a resampling-approximation permutation test does not require calculation of a hypergeometric probability value for each possible arrangement of cell frequencies.

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} .$$

If X and Y represent the row and column variables, respectively, there are $n(n-1)/2$ pairs of objects in the table that can be partitioned into five mutually exclusive exhaustive types: concordant pairs, discordant pairs, pairs tied on variable X but differing on variable Y, pairs tied on variable Y but differing on variable X, and pairs tied on both variable X and variable Y.

Concordant pairs (C) are pairs of objects that are ranked in the same order on both variable X and variable Y, given by

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \right) ;$$

discordant pairs (D) are pairs of objects that are ranked in one order on variable X and the reverse order on variable Y, given by

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \right) ;$$

pairs of objects tied on variable X but differing on variable Y (T_x) are given by

$$T_x = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \left(\sum_{k=j+1}^{c} n_{ik} \right) ;$$

pairs of objects tied on variable Y but differing on variable X (T_y) are given by

$$T_y = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \left(\sum_{k=i+1}^{r} n_{kj} \right) ;$$

and pairs of objects tied on both variable X and variable Y are given by

$$T_{xy} = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \left(n_{ij} - 1 \right) .$$

Given C, D, T_x, T_y, and n, then

$$\tau_a = \frac{2(C-D)}{n(n-1)} ,$$

$$\tau_b = \frac{C-D}{\sqrt{(C+D+T_x)(C+D+T_y)}} ,$$

$$\tau_c = \frac{(2m)(C - D)}{n^2(m - 1)},$$

where $m = \min(r, c)$,

$$\gamma = \frac{C - D}{C + D},$$

$$d_{yx} = \frac{C - D}{C + D + T_y},$$

and

$$d_{xy} = \frac{C - D}{C + D + T_x}.$$

Berry et al. illustrated exact and resampling-approximation permutation tests for τ_a, τ_b, τ_c, γ, d_{yx}, and d_{xy} with a variety of sparse data sets. The difference between the exact permutation tests and the resampling-approximation permutation tests is important to note. In the case of an exact permutation test of an $r \times c$ contingency table, it is necessary to calculate the selected measure of ordinal association for the observed cell frequencies and exhaustively enumerate all possible arrangements of the n objects in the rc cells, given fixed marginal frequency totals. Then, for each arrangement of cell frequencies, the measure of ordinal association, $T = \tau_a, \tau_b, \tau_c, \gamma, d_{yx}$, or d_{xy}, and the exact hypergeometric probability under the null hypothesis, $p(n_{ij}|n_{i.}, n_{.j})$, are calculated, where the hypergeometric probability value is given by

$$p(n_{ij}|n_{i.}, n_{.j}) = \frac{\left(\prod_{i=1}^{r} n_{i.}!\right)\left(\prod_{j=1}^{c} n_{.j}!\right)}{n!\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!},$$

$n_{i.}$ is the ith of r row marginal frequency totals, and $n_{.j}$ is the jth of c column marginal frequency totals. If T_{o} denotes the value of the observed test statistic, the exact one-sided upper- and lower-tail probability values of T_{o} are the sums of the $p(n_{ij}|n_{i.}, n_{.j})$ values associated with the T values computed on all possible arrangements of cell frequencies that are equal to or greater than T_{o} when T_{o} is positive, and equal to or less than T_{o} when T_{o} is negative, respectively.

By comparison, when the number of possible arrangements of cell frequencies is very large, exact permutation tests are impractical and resampling-approximation permutation methods become necessary, wherein a random sample of all possible arrangements of cell frequencies, drawn with replacement with fixed marginal frequency totals, is examined. The resampling-approximation one-sided upper- and lower-tail probability values of T_{o} are the proportions of the T values computed on the resampled arrangements of cell frequencies that are equal to or greater than T_{o} when T_{o} is positive and equal to or less than T_{o} when T_{o} is negative, respectively. It is readily apparent why resampling-approximation permutation methods are so efficient. First, only a random sample of all possible arrangements of cell frequencies needs to be examined. Second, and most importantly, resampling-approximation probability values are based simply on counting the number of T values equal to or more extreme than T_{o}, while exact permutation probability values require computation of the hypergeometric probability value, $p(n_{ij}|n_{i.}, n_{.j})$, for each arrangement of cell frequencies.

Comparisons of asymptotic, exact, and resampling-approximation one-sided probability values, where the resampling-approximation probability values were based on 1,000,000 random arrangements of the cell frequencies, demonstrated the advantages of exact and resampling permutation methods over asymptotic methods for the six statistics computed on sparse $r \times c$ contingency tables. Berry et al. concluded that the permutation methods utilized could be easily adapted to other measures of ordered association, non-sparse contingency tables, and two-sided probability values.

6.11 Resampling for Multi-Way Contingency Tables

Boyett in 1979 [199] and Patefield in 1981 [1089] developed resampling algorithms for $r \times c$ contingency tables (qq.v. pages 271 and 281). Both algorithms enumerated a subset of all possible two-way contingency tables from an observed contingency table with fixed marginal frequency totals. In 2002 Mielke and Berry published an article on testing for categorical independence in large sparse r-way contingency tables [962] and in 2007 Mielke, Berry, and Johnston presented a resampling algorithm for the enumeration of a subset of all possible r-way contingency tables with fixed marginal frequency totals [975]. To simplify presentation, the description here is restricted to a three-way contingency table and the example to a two-way contingency table, but the original algorithm provided corresponding analyses for any r-way contingency table with an integral value of $r \geq 2$. A moment-approximation permutation procedure based on the hypergeometric distribution for m-way contingency tables (q.v. page 298) was published in 1988 by Mielke and Berry [948]; see also [965, Sects. 7.1 and 7.3].

6.11.1 Description

Consider an $r \times c \times s$ contingency table with row marginals R_i, $i = 1, \ldots, r$, column marginals C_j, $j = 1, \ldots, c$, slice marginals S_k, $k = 1, \ldots, s$, and let the total number of objects classified be given by

$$N = \sum_{i=1}^{r} R_i = \sum_{j=1}^{c} C_j = \sum_{k=1}^{s} S_k .$$

In addition, set cell frequencies n_{ijk} equal to zero for $i = 1, \ldots, r$, $j = 1, \ldots, c$, and $k = 1, \ldots, s$, and set row, column, and slice counters, I, J, and K, respectively, equal to zero. Also note that all marginal frequency totals, R_i, $i = 1, \ldots, r$, C_j, $j = 1, \ldots, c$, S_k, $k = 1, \ldots, s$, and N are obtained from the observed contingency table and are fixed for all resamplings.

Calculate the cumulative row, column, and slice marginal proportions, PR_i, PC_j, and PS_k, respectively, for $i = 1, \ldots, r$, $j = 1, \ldots, c$, and $k = 1, \ldots, s$, where

$$PR_1 = R_1/N, \quad PR_i = PR_{i-1} + R_i/N, \quad \text{for } i = 2, \ldots, r,$$

$$PC_1 = C_1/N, \quad PC_j = PC_{j-1} + C_j/N, \quad \text{for } j = 2, \ldots, c,$$

and

$$PS_1 = S_1/N, \quad PS_k = PS_{k-1} + S_k/N, \quad \text{for } k = 2, \ldots, s.$$

Generate uniform pseudorandom numbers U_r, U_c, and U_s over $[0, 1)$ for the rows, columns, and slices, respectively, and set row, column, and slice indices $i = j = k = 1$, respectively. If $U_r \leq PR_i$, row counter $I = i$ and row marginal frequency total $R_i = R_i - 1$; if $U_c \leq PC_j$, column counter $J = j$ and column marginal frequency total $C_j = C_j - 1$; and if $U_s \leq PS_k$, slice counter $K = k$ and slice marginal frequency total $S_k = S_k - 1$. Finally, $N = N - 1$ and $n_{IJK} = n_{IJK} + 1$. The process is continued with new values of U_r, U_c, and U_s and terminated when $N = 0$.

6.11.2 An Example Analysis

For an example, consider a two-way contingency table with $r = c = 3$ as illustrated in Subtable A on the left side of Table 6.1. Subtable A has row marginal frequency totals of $R_1 = 7$, $R_2 = 6$, and $R_3 = 3$; column marginal frequency totals of $C_1 = 3$, $C_2 = 5$, and $C_3 = 8$; a total frequency of $N = 16$; and cell frequencies of $n_{ij} = 0$ for $i, j = 1, \ldots, 3$. The cumulative row proportions are $PR_1 = 7/16 = 0.44$, $PR_2 = (7 + 6)/16 = 0.81$, and $PR_3 = (7 + 6 + 3)/16 = 1.00$, and the cumulative column proportions are $PC_1 = 3/16 = 0.19$, $PC_2 = (3 + 5)/16 = 0.50$, and $PC_3 = (3 + 5 + 8)/16 = 1.00$.

Suppose that the row pseudorandom number is $U_r = 0.86$, which is greater than $PR_1 = 0.44$ and $PR_2 = 0.81$, but is less than or equal to $PR_3 = 1.00$. Thus, $I = i = 3$. Correspondingly, suppose that the column pseudorandom number is $U_c = 0.04$, which is less than or equal to $PC_1 = 0.19$. Thus, $J = j = 1$. Then $n_{IJ} = n_{3,1} + 1 = 0 + 1 = 1$ as illustrated in Subtable B on the right side of

Table 6.1 Example data set with $r = c = 3$, $N = 16$, and $n_{ij} = 0$ for $i, j = 1, \ldots, 3$ with marginal frequency totals R_i, $i = 1, \ldots, 3$ and C_j, $j = 1, \ldots, 3$ based on the fixed marginal frequency totals of an observed 3 × 3 contingency table

	Table A			R_i	PR_i	Table B			R_i	PR_i
	0	0	0	7	0.44	0	0	0	7	0.47
	0	0	0	6	0.81	0	0	0	6	0.87
	0	0	0	3	1.00	1	0	0	2	1.00
C_j	3	5	8	16		2	5	8	15	
PC_j	0.19	0.50	1.00			0.13	0.47	1.00		

Table 6.1. Subsequently, the row and column marginals R_3 and C_1 are decreased by 1, i.e., $R_3 - 1 = 3 - 1 = 2$ and $C_1 - 1 = 3 - 1 = 2$. The total frequency N is also decreased by 1, i.e., $N - 1 = 16 - 1 = 15$. New row and column cumulative proportions are then calculated, in this case resulting in the PR_i and PC_j values in Subtable B, $i, j = 1, \ldots, 3$: $PR_1 = 7/15 = 0.47$, $PR_2 = (7 + 6)/15 = 0.87$, $PR_3 = (7 + 6 + 2)/15 = 1.00$, $PC_1 = 2/15 = 0.13$, $PC_2 = (2 + 5)/15 = 0.47$, and $PC_3 = (2 + 5 + 8)/15 = 1.00$. Next, two new pseudorandom numbers are generated and the process repeated until $N = 0$.

6.12 Mielke–Berry and a Multivariate Similarity Test

In many areas of research it is often necessary to assess the similarity between multivariate measurements of corresponding unordered categories from two populations. In 2007 Mielke and Berry published a multivariate permutation test of similarity between two populations with corresponding unordered disjoint categories [966].

Following the notation of Mielke and Berry, consider two samples consisting of M and N objects in g unordered disjoint categories in which $m_i > 0$ and $n_i > 0$ are the number of objects in the ith of the g categories, for $i = 1, \ldots, g$; thus,

$$M = \sum_{i=1}^{g} m_i \quad \text{and} \quad N = \sum_{i=1}^{g} n_i \,.$$

Also consider that r distinct multivariate measurements may be associated with each object. Let $x_I = (x_{I1}, \ldots, x_{Ir})$ denote the row vector of r measurements for the Ith of M objects in Sample 1, let $y_J = (y_{J1}, \ldots, y_{Jr})$ denote the row vector of r measurements for the Jth of N objects in Sample 2, and assume that the observed M and N objects in Samples 1 and 2, respectively, are ordered so that the objects occur in the g categories according to the respective ordered category size structures $(m_1, \ldots, m_g)$ and $(n_1, \ldots, n_g)$ [966].

Let

$$s_i = \sum_{j=1}^{i} m_j \quad \text{and} \quad t_i = \sum_{j=1}^{i} n_j$$

for $i = 1, \ldots, g$. Also, let $s_0 = t_0 = 0$ and note that $s_g = M$ and $t_g = N$. If $\Delta_{I,J}$ is the r-dimensional Euclidean distance between the Ith and Jth objects in Samples 1 and 2, respectively, then

$$\Delta_{I,J} = \left[\sum_{k=1}^{r} \left(x_{Ik} - y_{Jk} \right)^2 \right]^{1/2} .$$

The average Euclidean distance between Sample 1 and Sample 2 objects in the ith category is given by

$$d_i = \frac{1}{m_i \, n_i} \sum_{I=s_{i-1}+1}^{s_i} \; \sum_{J=t_{i-1}+1}^{t_i} \Delta_{I,J}$$

for $i = 1, \ldots, g$. Then the two-sample multivariate permutation similarity comparison statistic is given by

$$W = \sum_{i=1}^{g} C_i d_i ,$$

where

$$C_i = \frac{(m_i n_i)^{1/2}}{\displaystyle\sum_{j=1}^{g} (m_j n_j)^{1/2}} .$$

As Mielke and Berry explained, the null hypothesis (H_0) posits that each of the $M!N!$ possible orderings of the M and N objects in Samples 1 and 2 is equally likely to occur. Thus, if Samples 1 and 2 are similar, the anticipated observed value of W will be smaller than expected under H_0. If W_o is the observed value of W, then the exact P value under H_0 is given by

$$P(W \leq W_\text{o}) .$$

If a random sample of L values of W is denoted by $W_1, \ldots, W_L$, then the approximate resampling P value associated with W_o is given by

$$P = \frac{1}{L} \sum_{i=1}^{L} \Psi_i ,$$

where

$$\Psi_i = \begin{cases} 1 & \text{if } W_i \leq W_\text{o}, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, Mielke and Berry provided an alternative asymptotic approximate normal P value given by $P(Z \leq Z_\text{o})$, where

$$Z_\text{o} = \frac{W_\text{o} - E[W]}{\hat{\sigma}_W},$$

the exact mean of W under H_0 is given by

$$E[W] = \frac{1}{MN} \sum_{I=1}^{M} \sum_{J=1}^{N} \Delta_{I,J},$$

an estimate of the standard deviation of W, σ_W, obtained from the resampling of the L values of W under H_0 is given by

$$\hat{\sigma}_W = \left[\frac{1}{L} \sum_{i=1}^{L} \Big(W_i - E[W] \Big)^2 \right]^{1/2},$$

and Z is a $N(0,1)$ random variable.

6.13 Cohen's Weighted Kappa with Multiple Raters

In 2008 Mielke, Berry, and Johnston utilized a resampling algorithm for r-way contingency tables (q.v. page 387) to analyze Cohen's weighted kappa with multiple raters [976]. The analysis of agreement using weighted kappa for multiple raters had long puzzled researchers. The usual procedure had been to examine all possible pairs of raters, akin to multiple two-sample t tests instead of a one-way analysis of variance F test; see for example, articles by Epstein, Dalinka, Kaplan, Aronchick, Marinelli, and Kundel [412], Herman, Khan, Kallman, Rojas, Carmody, and Bodenheimer [612], Kramer and Feinstein [769], Kundel and Polansky [781], Schouten [1234–1236], and Taplin, Rutter, Elmore, Seger, White, and Brenner [1340]. However, this approach made it impossible to provide an overall probability value because the pairwise comparisons were not orthogonal [976,977]. To simplify presentation, the description here is restricted to a three-way contingency table; in 2009 Mielke, Berry, and Johnston provided an example analysis of a four-way contingency table in *International Journal of Management* [977].

Consider $m = 3$ raters who independently classify n objects into $r = 5$ disjoint ordered categories. Conceptualize the classification as an $r \times r \times r$ contingency table with $r = 5$ rows, $r = 5$ columns, and $r = 5$ slices. Let n_{ijk}, R_i, C_j, and S_k denote the cell frequencies and the row, column, and slice marginal frequency totals for $i, j, k = 1, \ldots, r$, respectively, and let the frequency total for all r^3 cells be given by

$$N = \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} n_{ijk} .$$

Then, Cohen's weighted kappa test statistic for a three-way contingency table is given by

$$\hat{\kappa} = \frac{N^2 \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk}\, n_{ijk}}{\displaystyle\sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk}\, R_i\, C_j\, S_k} ,$$

where w_{ijk} are the weights assigned to each cell for $i, j, k = 1, \ldots, r$.

Employing the resampling algorithm for r-way contingency tables (q.v. page 387), where each dimension of the table represents a rater, Mielke, Berry, and Johnston generated r-way contingency tables with fixed marginal frequency totals for $L = 1{,}000{,}000$ random resamplings in an example analysis. If $\hat{\kappa}_{\text{o}}$ denotes the observed value of $\hat{\kappa}$, the resampling-approximation probability value for the observed kappa is given by

$$\hat{P}\,(\hat{\kappa}_{\text{o}}) = \frac{1}{L} \sum_{i=1}^{L} \Psi_i\,(\hat{\kappa}) ,$$

where

$$\Psi_i\,(\hat{\kappa}) = \begin{cases} 1 & \text{if } \hat{\kappa} \geq \hat{\kappa}_{\text{o}}, \\ 0 & \text{otherwise.} \end{cases}$$

The r-way resampling-approximation procedure yielded overall probability values that were much superior to probability values obtained from pairwise comparisons of raters [976].

In 2008, Berry, Johnston, and Mielke compared six procedures for unweighted kappa, weighted kappa with linear weighting,[17] and weighted kappa with quadratic weighting: exact variance, r-way resampling, intraclass correlation, randomized

[17] Linear weighting was first suggested by Cicchetti and Allison [255]. In 2008 Vanbelle and Albert demonstrated that weighted kappa for $m = 2$ independent raters and $r \geq 3$ ordered categories is equivalent to deriving the weighted kappa coefficient from unweighted kappa values computed on $r - 1$ embedded 2×2 classification tables, given linear weighting [1393]. In 2009 Mielke and Berry generalized the results of Vanbelle and Albert to $m \geq 2$ independent raters [967].

blocks, resampling blocks, and exact permutation analyses [115]. They concluded that the r-way contingency table resampling procedure provided the best estimates of the exact probabilities for symmetric unweighted, linear-weighted, and quadratic-weighted kappa. In 2009 Mielke, Berry, and Johnston demonstrated that for small samples, the r-way resampling-approximation procedure yielded more accurate results than the exact variance procedure advocated by Everitt [415], but noted that the advantage was mitigated with large samples as they both yielded unweighted and weighted kappa statistics that were asymptotically distributed as $N(0, 1)$ under the null hypothesis [977].

6.14 Exact Variance of Weighted Kappa

The determination of an accurate approximation to the exact variance of weighted kappa was a long-standing problem for many years; see for example, two papers by Hubert in 1977 and 1978 [667, 668]. Moreover, different weighting schemes oftentimes required different approximations to the exact variance [264]. Everitt had provided the exact variance of weighted kappa suitable for any weighting scheme in 1968, but found the expression too complicated for routine use [469, p. 323]. Consequently, a number of researchers attempted to provide estimates of the exact variance; see for example, Cohen in 1968 [264]; Everitt in 1968 [415]; Fleiss, Cohen, and Everitt in 1969 [469][18]; Cicchetti and Fleiss in 1977 [257]; Fleiss and Cicchetti in 1978 [468]; Hubert in 1978 [668]; Fleiss in 1981 [467]; Kramer and Feinstein in 1981 [769]; Banerjee, Capozzoli, McSweeney, and Sinha in 1999 [60]; Kingman in 2002 [757]; Ludbrook in 2002 [852]; Perkins and Becker in 2002 [1118]; Fleiss, Levin, and Paik in 2003 [470]; Kundel and Polansky in 2003 [781]; and Schuster in 2004 [1239].

As Hubert noted in 1978, in many cases, such as in articles by Fleiss, Cohen, and Everitt in 1969 [469] and Cohen in 1972 [265], the large sample variance of weighted kappa was derived under the assumption that only the total sample size was fixed [668]. However, for many of the historically important applications of an index of nominal-level agreement Hubert argued that it is more appropriate to assume fixed marginal frequency totals [668, p. 184]. In 2005 Mielke, Berry, and Johnston reformulated the exact variance formula presented by Everitt in 1968 into a form conducive to computation, and in 2007 they extended Everitt's formula for $m = 2$ raters to $m \geq 2$ raters and provided an efficient computer algorithm [973, 974]. Although the algorithm described by Mielke et al. is appropriate for any

[18] In 1972 Cohen admitted that the formulae for the approximate variance of weighted kappa given by Cohen in 1968 [264] and by Everitt in 1968 [415] were both incorrect [265, p. 64], but that the formula given by Fleiss, Cohen, and Everitt in 1969 [469] was, in fact, correct. This latter statement turned out to be incorrect.

number of raters, the discussion here is restricted to $m = 3$ raters for simplicity and clarity.[19]

As described previously, for $m = 3$ raters and N objects cross-classified into an $r \times r \times r$ contingency table composed of an ordered categorical variable with r rows, r columns, and r slices, let n_{ijk}, w_{ijk}, R_i, C_j, and S_k for $i, j, k = 1, \ldots, r$ denote the cell frequencies, cell weights, row marginal frequency totals, column marginal frequency totals, and slice marginal frequency totals, respectively, where

$$R_i = \sum_{j=1}^{r} \sum_{k=1}^{r} n_{ijk}, \quad C_j = \sum_{i=1}^{r} \sum_{k=1}^{r} n_{ijk}, \quad \text{and} \quad S_k = \sum_{i=1}^{r} \sum_{j=1}^{r} n_{ijk},$$

and the frequency total is given by

$$N = \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} n_{ijk}.$$

Given fixed row, column, and slice marginal frequency totals, the weighted kappa test statistic for $m = 3$ raters is defined as

$$\hat{\kappa} = 1 - \frac{N^2 \displaystyle\sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk} n_{ijk}}{\displaystyle\sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk} R_i C_j S_k}$$

and the exact mean and variance of $\hat{\kappa}$ under the null hypothesis are given by $\mu_{\hat{\kappa}} = 0$ and

$$\sigma_{\hat{\kappa}}^2 = \frac{N^4 \operatorname{E}\left[\left(\displaystyle\sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk} n_{ijk}\right)^2\right]}{\left(\displaystyle\sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk} R_i C_j S_k\right)^2} - 1,$$

respectively. The reformulation of the exact variance of $\hat{\kappa}$ is based on

$$\sigma_{\hat{\kappa}}^2 = \frac{W N^2}{G^2 (N-1)^2} - 1,$$

[19] In 2009 Mielke, Berry, and Johnston provided an example analysis based on $m = 4$ independent raters for both unweighted and weighted kappa that was published in *International Journal of Management* [977].

where

$$G = \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk} R_i C_j S_k ,$$

$$W = Y + Y_1 + Y_2 + Y_3 + Y_{12} + Y_{13} + Y_{23} + Y_{123},$$

$$Y = \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk}^2 R_i C_j S_k \Big[(R_i - 1)(C_j - 1)(S_k - 1) + (N - 1)^2 \Big],$$

$$Y_1 = \sum_{i \neq i'}^{r} \sum_{j=1}^{r} \sum_{k=1}^{r} w_{ijk} w_{i'jk} R_i C_j S_k (C_j - 1)(S_k - 1) R_{i'},$$

$$Y_2 = \sum_{i=1}^{r} \sum_{j \neq j'}^{r} \sum_{k=1}^{r} w_{ijk} w_{ij'k} R_i C_j S_k (R_i - 1)(S_k - 1) C_{j'},$$

$$Y_3 = \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k \neq k'}^{r} w_{ijk} w_{ijk'} R_i C_j S_k (R_i - 1)(C_j - 1) S_{k'},$$

$$Y_{12} = \sum_{i \neq i'}^{r} \sum_{j \neq j'}^{r} \sum_{k=1}^{r} w_{ijk} w_{i'j'k} R_i C_j S_k (S_k - 1) R_{i'} C_{j'},$$

$$Y_{13} = \sum_{i \neq i'}^{r} \sum_{j=1}^{r} \sum_{k \neq k'}^{r} w_{ijk} w_{i'jk'} R_i C_j S_k (C_j - 1) R_{i'} S_{k'},$$

$$Y_{23} = \sum_{i=1}^{r} \sum_{j \neq j'}^{r} \sum_{k \neq k'}^{r} w_{ijk} w_{ij'k'} R_i C_j S_k (R_i - 1) C_{j'} S_{k'},$$

$$Y_{123} = \sum_{i \neq i'}^{r} \sum_{j \neq j'}^{r} \sum_{k \neq k'}^{r} w_{ijk} w_{i'j'k'} R_i C_j S_k R_{i'} C_{j'} S_{k'},$$

and $\sum_{i \neq i'}^{r}$ denotes the sum of all terms with $i \neq i'$ for a specified category [974]. Then,

$$Z = \frac{\hat{\kappa} - \mu_{\hat{\kappa}}}{\sigma_{\hat{\kappa}}}$$

approaches the $N(0, 1)$ distribution as $N \rightarrow \infty$ under the null hypothesis with fixed positive marginal proportions. Let $\hat{\kappa}_{\text{o}}$ denote the observed value of $\hat{\kappa}$, then the

Table 6.2 Eighty-five article reviews by $m = 3$ reviewers with $r = 3$ categories: reject, revise and resubmit, and accept

		Column		
Slice	Row	Reject	Revise	Accept
Reject	Reject	7 (0)[a]	4 (2)	0 (4)
	Revise	5 (2)	2 (2)	1 (4)
	Accept	1 (4)	3 (4)	2 (4)
Revise	Reject	4 (2)	5 (2)	3 (4)
	Revise	5 (2)	8 (0)	3 (2)
	Accept	0 (4)	7 (2)	2 (2)
Accept	Reject	3 (4)	4 (4)	4 (4)
	Revise	3 (4)	2 (2)	2 (2)
	Accept	1 (4)	2 (2)	2 (0)

[a] Numbers in parentheses are linear weights for the weighted kappa statistic

approximate probability (P) value under the null hypothesis is given by $P(Z \geq Z_{\mathrm{o}})$, where Z is a $N(0, 1)$ random variable and

$$Z_{\mathrm{o}} = \frac{\hat{\kappa}_{\mathrm{o}}}{\sigma_{\hat{\kappa}}}.$$

6.14.1 An Example Analysis

For an example analysis, consider $m = 3$ independent reviewers, each of which has been presented with $N = 85$ articles to review for a scholarly journal over a 5-year period. Each reviewer submits one of three possible recommendations for each article: reject, revise and resubmit, or accept. Table 6.2 contains the $r^3 = 27$ cross-classified recommendations and the corresponding linear weights given in parentheses, where

$$w_{ijk} = |i - j| + |i - k| + |j - k|$$

for $i, j, k = 1, \ldots, r$. For the observed data and associated linear weights in Table 6.2, $\hat{\kappa}_{\mathrm{o}} = +0.1130$, $\sigma^2_{\hat{\kappa}} = 0.2292 \times 10^{-2}$, $Z_{\mathrm{o}} = +2.3603$, and the approximate $N(0, 1)$ one-sided upper-tail probability value is 0.9131×10^{-2}. For comparison, a resampling analysis based on $L = 1{,}000{,}000$ random arrangements of the cell frequencies in Table 6.2 yielded $\hat{\kappa}_{\mathrm{o}} = +0.1130$ and an approximate one-sided upper-tail probability value of 0.1474×10^{-1}. In 1960 Cohen stated that, although kappa can assume a negative value when the magnitude of agreement is less than chance expectancy, negative values are to be interpreted as zero agreement since they indicate a general lack of agreement among the m raters. Consequently, only the right tail of the distribution is to be considered with a one-sided upper-tail probability value [1261, p. 546].

Table 6.3 Nine article reviews by $m = 3$ reviewers with $r = 3$ categories: reject, revise and resubmit, and accept

		Column		
Slice	Row	Reject	Revise	Accept
Reject	Reject	1 (0)[a]	0 (2)	0 (4)
	Revise	1 (2)	0 (2)	0 (4)
	Accept	0 (4)	0 (4)	0 (4)
Revise	Reject	1 (2)	0 (2)	0 (4)
	Revise	0 (2)	1 (0)	0 (2)
	Accept	0 (4)	0 (2)	1 (2)
Accept	Reject	0 (4)	0 (4)	1 (4)
	Revise	0 (4)	1 (2)	0 (2)
	Accept	0 (4)	0 (2)	2 (0)

[a] Numbers in parentheses are linear weights for the weighted kappa statistic

While the probability values based on the exact variance and the resampling analysis are close to each other, i.e., 0.9131×10^{-2} and 0.1474×10^{-1}, respectively, this is due to the relatively large sample size of $N = 85$ and a relatively low resampling skewness value of $\gamma_{\hat{\kappa}} = 0.1421$. In contrast, consider a second example with $m = 3$ independent reviewers presented with only $N = 9$ articles to review. Table 6.3 contains the $r^3 = 27$ cross-classified recommendations with the corresponding linear weights given in parentheses. For the observed data and associated linear weights in Table 6.3, $\hat{\kappa}_{\text{o}} = +0.5091$, $\sigma^2_{\hat{\kappa}} = 0.2583 \times 10^{-1}$, $Z_{\text{o}} = +3.1678$, and the approximate $N(0, 1)$ one-sided upper-tail probability value is 0.7679×10^{-3}. For comparison, a resampling analysis based on $L = 1{,}000{,}000$ random arrangements of the data in Table 6.3 yielded $\hat{\kappa}_{\text{o}} = +0.5091$ and an approximate one-sided upper-tail probability value of 0.4385×10^{-2}. In this case, with a small sample size of $N = 9$ and a relatively high resampling skewness of $\gamma_{\hat{\kappa}} = 0.4435$, the exact-variance and resampling probability values of 0.7679×10^{-3} and 0.4385×10^{-2}, respectively, are quite different. While the resampling method is preferred, the normal approximation based on the exact variance provides improved results with increasing sample size due to the asymptotic normality.

6.15 Campbell and Two-by-Two Contingency Tables

In 2007 Campbell published a lengthy article in *Statistics in Medicine* on the chi-squared and Fisher–Irwin tests of 2×2 contingency tables with small sample sizes [239]. Campbell noted that this had been an issue of some interest for over 100 years and dozens of research papers had been devoted to the topic [239, p. 3,662].

Fig. 6.2 Notation for a 2 × 2 contingency table as used by Campbell [239, p. 3662]

	B	not-B	Total
A	a	b	m
not-A	c	d	n
Total	r	s	N

Given a 2 × 2 contingency table such as depicted in Fig. 6.2, Campbell distinguished among three distinct research designs.[20] In the first research design, termed a *comparative trial* design, there are two populations and samples of sizes, m and n, are taken from the first and second populations, respectively. In this design, the research question is whether the proportions of B in the two populations are the same and the row marginal frequency totals, m and n of variable A, are assumed to be fixed. In the second research design, termed a *cross-sectional* design, a single sample of size N is drawn from a population and each member of the sample is classified according to two binary variables, A and B. In this design, the row totals are not determined by the investigator and no marginal frequency totals are assumed to be fixed. In the third research design, termed an *independence trial* design, both sets of marginal frequency totals for variables A and B are assumed to be fixed by the investigator.

Next, Campbell detailed three versions of the chi-squared test for 2 × 2 contingency tables.[21] He noted that in the original version of the chi-squared test, due to K. Pearson in 1900 [1107] and R.A. Fisher in 1922 [446], the value of the expression

$$\frac{N\,(ad - bc)^2}{mnrs}$$

is evaluated with the chi-squared distribution with one degree of freedom.[22] In 1934 Yates (q.v. page 37) recommended an adjustment to the expression that came to be known as the Yates' continuity correction, where

$$\frac{N\left(|ad - bc| - \frac{N}{2}\right)^2}{mnrs}$$

is evaluated with the chi-squared distribution with one degree of freedom [1472].[23] In 1947 E.S. Pearson recommended a third version of the chi-squared

[20] The three research designs were first described by Barnard in an article in *Biometrika* in 1947 [67] (q.v. page 130).

[21] For an excellent discussion of the chi-squared test for 2 × 2 contingency tables, see a 1990 article by John Richardson in *British Journal of Mathematical and Statistical Psychology* [1170].

[22] Karl Pearson had miscalculated the degrees of freedom in 1900 and it was corrected by Fisher in 1922, which did little to improve their antagonistic relationship.

[23] As Egon Pearson noted in 1947, the correction for continuity utilized by Yates in 1934 was not new at the time, having been used by statisticians for many years prior when employing a normal or skew curve to give the sum of terms of a binomial or hypergeometric series [1095, p. 147].

test, where the expression

$$\frac{(N-1)(ad-bc)^2}{mnrs}$$

is evaluated with the chi-squared distribution with one degree of freedom [1095].[24] Consider the numbers of B and not-B in the two samples in Fig. 6.2 and consider the research question that the proportion of B in both samples is the same, with the common proportion denoted by π. Then, as Campbell noted, while an unbiased estimate of π is r/N, an unbiased estimate of $\pi(1-\pi)$ is not $(r/N)(1-r/N)$, but is instead $(r/N)(1-r/N)N/(N-1)$ [239, p. 3663]; see also a discussion on this topic by Stuart, Ord, and Arnold in the sixth (1999) edition of Kendall's *Advanced Theory of Statistics* [1329, p. 17].

In the same manner as chi-squared, Campbell distinguished among four versions of the Fisher–Irwin exact probability test for 2 × 2 contingency tables. Originally developed by Fisher [446] and Irwin [674] as a one-sided test, the expression

$$\frac{m!\,n!\,r!\,s!}{N!\,a!\,b!\,c!\,d!}$$

provided the hypergeometric point-probability value of the observed table (q.v. page 48). Campbell noted that there are four versions of the Fisher–Irwin exact two-sided probability test.[25] The first version, advocated by Fisher, is to double the one-sided probability value, i.e., the doubling rule (q.v. page 51). The second version, advanced by Irwin, calculates the total probability of tables in either tail that are as likely as, or less likely than the observed probability value, i.e., Irwin's rule (q.v. page 51). The third version is a so-called "mid-*P*" test where only half the probability value of the observed contingency table is included in the one-sided probability value, then the one-sided probability value is doubled, as described by Berry and Armitage [107]. The fourth version is a mid-*P* test where the two-sided probability value is one-half the probability value of the observed contingency table plus the probability values of those contingency tables in either tail that are less than that of the observed contingency table, as described by Hirji, Tan, and Elashoff [632].

In a thorough examination and comparison of the three versions of the chi-squared test and the four versions of the Fisher–Irwin exact probability test over the three research designs (comparative, cross-sectional, and independence trials), Campbell concluded:]

1. For designs when both marginals are fixed (independence trial), Yates' correction for continuity is appropriate, but not for comparative trials or cross-sectional designs.

[24] On this topic, see articles by Barnard [65], Mielke and Berry [947,948], Richardson [1170,1171], Berry and Mielke [136], Upton [1385], and Schouten, Molenaar, van Strik, and Boomsma [1237].

[25] Campbell, unfortunately, neglected to note that the test was also independently developed by Yates in 1934 (q.v. page 43).

2. Where all expected cell frequencies are at least 1, use the "$N - 1$" chi-squared test recommended by E.S. Pearson.
3. Otherwise, analyze the data by the Fisher–Irwin exact probability test, with a two-sided test carried out by Irwin's rule.

In 2008 Martín Andrés commented on Campbell's paper in the same journal [897]. While the comments are interesting and informative, they focus on statistical matters unrelated to permutation methods. On this topic, see also a 2005 article by Martín Andrés, Sánchez Quevedo, Tapia García, and Silva Mato [903].

6.16 Permutation Tests and Robustness

Conventional statistical tests are hampered by a variety of assumptions that very often are not realistic, e.g., normally distributed populations.[26] As R.C. Geary famously put it in 1947, "[n]ormality is a myth; there never has, and never will be, a normal distribution" [501, p. 241]. The French physicist and Nobel laureate in physics, Gabriel Lippmann, once wrote in a letter to Henri Poincaré à propos the normal curve:

> Les expérimentateurs s'imaginent que c'est un théorèm de mathématiques, et les mathématiciens d'être un fait expérimental.
> Experimentalists think that it is a mathematical theorem, while mathematicians believe it to be an experimental effect.

(Lippman, quoted in D'Arcy Wentworth Thompson's *On Growth and Form* [1358, p. 121]). And in 1954 Bross pointed out that statistical methods "are based on certain assumptions—assumptions which not only can be wrong, but in many situations *are* wrong" [222, p. 815].[27,28] Others have empirically demonstrated the prevalence of highly-skewed and heavy-tailed distributions in a variety of academic disciplines; see for example, discussions by Schmidt and Johnson [1233], Bradley [202], Saal, Downey, and Lahey [1212], Bernardin and Beatty [104], Micceri [930], and Murphy and Cleveland [1016], the best known of which is Micceri's widely quoted 1989 article on "The unicorn, the normal curve, and other improbable creatures" in *Psychological Bulletin* [930].

For example, in 2012 O'Boyle and Aguinis published an article in *Personnel Psychology* in which they studied 633,263 people in five broad areas of human performance: 490,185 researchers who produced 943,224 publications across 54 academic disciplines between January 2000 and June 2009 with performance measured by number of publications; 17,750 individuals in the entertainment industry with performance measures based on the number of times an entertainer

[26] For reasons why the assumption of normality is critical in conventional statistical analyses, see a 2011 paper by Mordkoff [1006].

[27] Emphasis in the original.

[28] See also a short but comprehensive 2010 article on this topic by Tom Siegfried in *Science News* [1274].

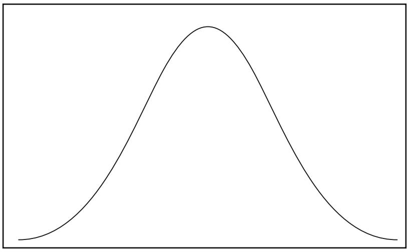


Fig. 6.3 A typical Gaussian curve.

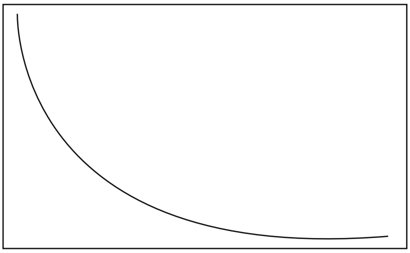


Fig. 6.4 A typical Paretian curve

received an award, nomination, or other indicator; 42,745 candidates running for political office in Australia, Canada, Denmark, Estonia, Finland, Holland, Ireland, New Zealand, the United Kingdom, and the United States, where performance was based on the number of times the candidate had been elected to a political office; 25,283 athletes in a variety of collegiate and professional sports, where performance was based on different positive performance criteria for different sports, e.g., home runs for baseball, number of wins for tennis and golf, and goals or points for soccer and hockey; and 57,300 athletes in a variety of collegiate and professional sports, where negative performance was based on different criteria for different sports, e.g., number of errors in baseball, incomplete passes in football, dropped passes in football, yellow cards in soccer, and so on [1044].

Using a chi-squared goodness-of-fit test as a measure of misfit, O'Boyle and Aguinis concluded that in all five studies the distribution of performance followed a Paretian (Pareto) power distribution more closely then a Gaussian (normal) distribution.[29] For example, in their research on academic publications, they found the average misfit for the Paretian distribution to be 23,888, while the misfit of the normal distribution was 44,199,201,241,681—a difference in favor of the Paretian distribution on the order of 1:1.9 billion [1044, p. 87]. Figure 6.3 depicts a typical Gaussian curve and Fig. 6.4 depicts a typical Paretian curve, similar to those described in O'Boyle and Aguinis [1044, p. 80].

[29] In 1907, the Italian economist and sociologist Vilfredo Pareto created a mathematical formula to describe the unequal distribution of wealth in Italy, observing that 20 % of the people owned 80 % of the wealth [1087]. This became known as the Pareto Principle or Pareto's Law. In general, the 80/20 rule has come to mean that in anything, a few (20 %) are vital and many (80 %) are trivial.

O'Boyle and Aguinis observed three important differences between the Gaussian and Paretian distributions in assessing individual performance. First, the Gaussian distributions underpredicted the likelihood of extreme events. Second, the Gaussian distributions assume that the mean and standard deviation are stable. However, if the underlying distribution is Paretian instead of Gaussian, means and standard deviations are not stable and Gaussian-based point estimates and confidence intervals are biased; see also a 2009 article by Andriani and McKelvey on this topic in *Organization Science* [22]. Third, a key difference between Gaussian and Paretian distributions is scale invariance, i.e., the extent to which a measurement instrument generalizes across different cultures or populations. Scale invariance also refers to the distribution remaining constant whether one is looking at the whole population or only the top performers [1044]. O'Boyle and Aguinis found the Paretian distribution to possess scale invariance, while the Gaussian distribution did not.

O'Boyle and Aguinis noted that the assumption of normality, like random sampling, belongs to the class of "received doctrines" that are

> [t]aught in undergraduate and graduate classes, enforced by gatekeepers (e.g., grant panels, reviewers, editors, dissertation committee members), discussed among colleagues, and otherwise passed along among pliers of the trade far and wide and from generation to generation [795, p. 281].

Finally, O'Boyle and Aguinis cautioned that "assuming normality . . . can lead to misspecified theories and misleading practices" [1044, p. 116].

6.16.1 Robustness and Rank-Order Statistics

The inclusion of extreme values (outliers) in data sets, both experimental and non-experimental, is common. This often necessitates an adjustment to the data to fit the assumptions of the statistics employed. While truncation or trimming of extreme values is not unusual, by far the most common methodological strategy is to convert the original measurements to rank-order statistics. Writing in 1954, Bross labeled rank-order statistics as "mutations" of conventional statistics [222, p. 815]. The replacement of numerical variates with their corresponding rank-order statistics to avoid the assumption of normality has a long history. Notable in this regard was the work of Spearman in 1904, 1906, and 1910 [1300–1302], Karl Pearson in 1907 [1109], Hotelling and Pabst in 1936 [653], Friedman in 1937 [485], Kendall in 1938 [728], Wilcoxon in 1945 [1453], Mann and Whitney in 1947 [880], and Kendall in 1948 [734]. Bross observed that rank transformations were first suggested by Spearman in 1904, but were so criticized by mathematical statisticians that no one dared to use them for 25 years [222]. Currently, many researchers consider rank-order statistics to be passé and prefer permutation tests utilizing the original numerical values over rank alternatives; see for example, a 2002 article by Berry, Mielke, and Mielke in *Psychological Reports* [162].

Although the advantages of rank-order statistics are well known [779], it is generally recognized that there is a loss of information when substituting numerical

Fig. 6.5 World education scores and rankings in mathematics for 12 selected countries

Country	Score	Rank
Shanghai–China	600	1
Singapore	562	2
Canada	527	3
Netherlands	526	4
Macao–China	525	5
Norway	498	6
France	497	7
Austria	496	8
Poland	495	9
Sweden	494	10
Czech Republic	493	11
Panama	360	12

values with rank-order statistics; see for example, articles by Feinstein [421], Friedman [485], Gebhard and Schmitz [502], Lehmann [815], Spearman [1301], Wald and Wolfowitz [1406, p. 387], and Still and White [1324]. In addition, because ranking methods are divorced from the original scale, they are not useful in the problem of estimation, as noted by Bross in 1954 [222]. Finally, the conversion of raw scores to ranks can make small differences between raw scores appear large and, conversely, make large differences between raw scores appear small. Figure 6.5 illustrates the potential for misleading information when raw score measurements are converted to rank-order statistics.

The data in Fig. 6.5 were extracted from the 2009 world education scores and rankings in mathematics and provided by the Programme for International Student Assessment (PISA) through the Organization for Economic Cooperation and Development (OECD) [1260, 1366]. Shanghai–China (specifically, the city of Shanghai in China) and the city-state of Singapore actually were ranked 1 and 2 in the world in 2009, respectively, but the difference in raw scores was $600-562 = 38$ points, which is a substantial difference considering that the range of scores for all 65 participating countries was only 270 points. Similarly, consider the raw-score and rank differences between the Czech Republic and Panama where the difference in raw scores in Fig. 6.5 was $493 - 360 = 133$ points, but the difference in ranks was only $12 - 11 = 1$. On the other hand, the raw score difference between Canada and the Czech Republic was $527 - 493 = 34$ points, 4 points less than the raw-score difference between Shanghai–China and Singapore, yet the difference in rankings was $11 - 3 = 8$ times the difference in ranks between Shanghai–China and Singapore, i.e., $2 - 1 = 1$. Thus, the conversion of raw-score measurements to ranks can both minimize large raw-score differences and magnify small raw score differences, with a consequent loss of information.

In this context, May and Hunter went so far as to label the practice of replacing observations with rank numbers as a "degrading of the original data" [908, p. 404], Arbuckle and Aiken bemoaned it as a "sacrifice of desirable qualities" [30, p. 381], Borgatta concluded that reality is distorted by assigning ranks and performing arithmetic operations on a set of numbers that is not isomorphic with the arithmetic system [182], and in 2008 Ludbrook noted that rank-order tests are the poor man's substitute for computer-intensive measures, concluding "I see no merit in using this class of test on interval-scale data" [853, p. 673]. As early as 1940 Kendall and Babington Smith acknowledged that "the use of ranking . . . destroys what may be valuable information" [741, p. 324], but also pointed out that the ranking method suffers from a serious drawback when the quality considered is known not to be representable by a linear variable [741, p. 324]. In a strongly worded statement in 2000 in reference to converting raw scores to ranks for the Wilcoxon two-sample rank-sum test, Ludbrook and Dudley argued that "although the [Wilcoxon–Mann–Whitney] test was a brilliant invention by Frank Wilcoxon in the pre-computer era as a way of overcoming the computation difficulties of executing a permutation test for equality of means, it should have little relevance today" [857, p. 87].[30] F.N. David, in a review of Kendall's *Rank Correlation Methods* commented:

> [i]t is interesting to note in the univariate case . . . that while many order statistics have been proposed (all of which are easy to apply and interesting mathematically) . . . it is rare indeed to find the need to use them in practice. It is customary to twist the observations about and/or to make various assumptions in order that existing techniques may be applied. This, the writer would suggest, is because of the instinctive feeling that tests based on ranks cannot be very discriminating. If, on the other hand, we consider the bivariate case, the order statistics proposed by Spearman and latterly by Kendall are used fairly frequently with little thought of the undoubted loss of information which using them implies [319, p. 190].

6.16.2 Mielke et al. and Robustness

Lehmann [816, p. 93], following Huber [665], defined robustness as when small deviations from the model result in small changes in the performance of the procedure, and conversely, when large deviations from the model result in large changes in the performance of the procedure.

In 2011 Mielke, Berry, and Johnston considered the topic of robustness without rank-order statistics in a permutation context [978]. In this article they considered an alternative to conventional rank tests that was based on a Euclidean-distance analysis space. Let a distance function between objects I and J be denoted by $\Delta_{I,J}$. Then the distance function affiliated with classical tests such as Student's two-sample t test, the k-sample F test, and the Wilcoxon–Mann–Whitney two-sample rank-sum test is the squared Euclidean distance function given by $\Delta_{I,J} = (y_I - y_J)^2$. They argued that if the parametric assumption of normality is removed, as in a rank test, there is no theoretical justification for a distance function such as

[30] There is a counter argument, of course, that power may not be lost when converting raw scores to ranks, and may even be increased, depending on which assumptions are violated and in which manner; see for example articles by Blair and Higgins [168], Higgins and Blair [614], Good [530], Hodges and Lehmann [635], Keller-McNulty and Higgins, [714], Lehmann [815], and van den Brink and van den Brink [1389].

Table 6.4 Raw-score observed values for $g = 2$ groups with $n_1 = n_2 = 13$ objects randomly assigned to each group

Group 1			Group 2		
264.3	264.9	265.2	263.4	264.0	264.3
264.6	264.9	265.5	263.7	264.0	264.6
264.6	264.9	265.5	263.7	264.3	w
264.6	265.2		263.7	264.3	
264.9	265.2		264.0	264.3	

$\Delta_{I,J} = (y_I - y_J)^2$. They then considered the general class of distance functions given by $\Delta_{I,J} = |y_I - y_J|^v$, concluding that only $v = 1$, i.e., Euclidean distance, ensured that the analysis space corresponded to the data space; see also a 2010 article on this topic by Reiss, Stevens, Shehzad, Petkova, and Milham in *Biometrics* [1163].

One of the early critics of squaring differences was Charles Spearman [1302]. In 1906, in the context of measures of correlation, Spearman wrote that "squaring is . . . more likely to do harm than good" [1301, p. 100]. In 1910, writing about rank-order correlation, and responding to harsh criticism by Karl Pearson [1108, 1109], Spearman wrote that "squaring lays stress on the extreme discrepancies between the series compared" [1302, p. 284]. Noting that squaring is based on the assumption of a Gaussian distribution, Spearman observed:

> [t]he Gaussian assumption is only a mathematical make-shift; we may often conveniently enough reckon formulae from it; but in actual application, we should constantly bear in mind its real limitations [1302, p. 285].

Examples given by Mielke et al. in 2011 were based on two-sample tests and illustrated the advantages of utilizing a Euclidean-distance function ($v = 1$) over a squared Euclidean-distance function ($v = 2$) [978]. Table 6.4 lists data for two independent groups of sizes $n_1 = n_2 = 13$. While the $n_1 = 13$ data points in Group 1 are fixed, one value in Group 2, designated by w, is allowed to vary in order to determine its effect on the exact two-sided probability values. Table 6.5 lists ten values for w ranging from a low value of $w = 40$ up to a high value of $w = 988$, the exact two-sided probability values for the Fisher–Pitman two-sample permutation test with $v = 1$ and $v = 2$, the exact two-sided probability values for the permutation version of the Wilcoxon–Mann–Whitney two-sample rank-sum test that involves $\Delta_{I,J}$ with $v = 2$,[31] and the two-sided probability values for the classical Student two-sample t test, under the usual assumptions of normality and independence. Each of the exact two-sided probability values in Table 6.5 is based on

[31] The exact Fisher–Pitman two-sample permutation tests with $v = 1, 2$ and the exact permutation version of the Wilcoxon–Mann–Whitney two-sample rank-sum test are specific forms of MRPP (q.v. page 249) with $g = 2$ and $r = 1$ (q.v. page 256).

Table 6.5 Two-sided probability value comparisons for the exact Fisher–Pitman two-sample test with $v = 1$ and $v = 2$, the exact two-sample Wilcoxon–Mann–Whitney (WMW) rank-sum test, and the classical Student two-sample t test for the data listed in Table 6.4

	Exact permutation test			
	Fisher–Pitman test		WMW	Student
w	$v = 1$	$v = 2$	rank test	t test
40	4.038×10^{-6}	4.038×10^{-6}	4.038×10^{-6}	0.303
240	4.038×10^{-6}	4.038×10^{-6}	4.038×10^{-6}	0.148
258	4.038×10^{-6}	4.038×10^{-6}	4.038×10^{-6}	0.009
261	4.038×10^{-6}	4.038×10^{-6}	4.038×10^{-6}	2.646×10^{-4}
264	4.038×10^{-6}	4.038×10^{-6}	4.038×10^{-6}	5.837×10^{-7}
267	9.115×10^{-5}	0.016	1.765×10^{-4}	0.016
270	9.115×10^{-5}	0.473	1.765×10^{-4}	0.346
276	9.115×10^{-5}	1.000	1.765×10^{-4}	1.000
288	9.115×10^{-5}	1.000	1.765×10^{-4}	0.622
988	9.115×10^{-5}	1.000	1.765×10^{-4}	0.335

$$M = \frac{(n_1 + n_2)!}{n_1!\, n_2!} = \frac{(13 + 13)!}{13!\, 13!} = 10{,}400{,}600$$

equally-likely arrangements of the 26 data values given in Table 6.4, with w included. The two-sided probability values for the classical two-sample t test are based on Student's t distribution with $n_1 + n_2 - 2 = 13 + 13 - 2 = 24$ degrees of freedom.

As illustrated in Table 6.5, the exact two-sided probability values for the Fisher–Pitman two-sample permutation test with $v = 1$ are stable, consistent, and unaffected by the extreme values of w in either direction. Similar results are achieved with the Wilcoxon–Mann–Whitney two-sample rank-sum test [880, 1453]. In contrast, the exact two-sided probability values for the Fisher–Pitman two-sample permutation test with $v = 2$ range from 4.038×10^{-6} for small values of w up to 1.000 for large values of w, relative to the fixed values.

Finally, the two-sided probability values for the classical two-sample t test approach a common value as w becomes very small or very large, relative to the fixed values, and the classical t test is unable to detect the obvious differences in location between Groups 1 and 2. Mielke et al. concluded that the exact Fisher–Pitman two-sample permutation test with $v = 1$ was robust to extreme values without the use of rank-order statistics. In 2000 Mielke and Berry conducted similar comparisons of univariate MRPP with $v = 1$ and $v = 2$ where the same conclusions were reached, i.e., $v = 1$ was far more robust than $v = 2$ when extreme values were included [959, pp. 13–15].

Another property to be considered is that MRPP are median-based procedures when $v = 1$, whereas MRPP are mean-based procedures when $v = 2$ (q.v. page 249). For clarification, consider the pairwise sum of univariate ($r = 1$) symmetric distance functions given by

$$\sum_{i<j} \Delta(x_i, x_j) = \sum_{i<j} |x_i - x_j|^v,$$

where $x_1, \ldots, x_m$ are univariate response variables and $\sum_{i<j}$ is the sum over all i and j such that $1 \leq i < j \leq m$. Let $x_{1,m} \leq \cdots \leq x_{m,m}$ be the order statistics associated with $x_1, \ldots, x_m$. If $v = 1$, then the inequality given by

$$\sum_{i=1}^{m} |m - 2i + 1| |x_{i,m} - \theta| \geq \sum_{i<j} |x_i - x_j|$$

holds for all θ, and equality holds if θ is the median of $x_1, \ldots, x_m$. If $v = 2$, then the inequality given by

$$m \sum_{i=1}^{m} (x_i - \theta)^2 \geq \sum_{i<j} (x_i - x_j)^2$$

holds for all θ, and equality holds if and only if θ is the mean of $x_1, \ldots, x_m$. Since most statistical tests are based on means [873, p. 434] and means can be severely influenced by a relatively few extreme values, whereas medians are seldom similarly affected, this property further explains why MRPP based on $v = 1$ are immensely more robust than MRPP based on $v = 2$ (q.v. page 254). Specifically, $v = 2$ places more emphasis on the extremes of a distribution than on the mass, while $v = 1$ places more emphasis on the mass than on the extremes. With the exception of ordinal data, robustness with $v = 1$ is apparently as good as or better than using rank-order statistics for the same purpose. Furthermore, the common use of rank-order statistics with $v = 2$, e.g., the Wilcoxon–Mann–Whitney and Kruskal–Wallis rank-sum tests, are also further improved when based on $v = 1$; see [978].

Consider a second example based on two matched groups (q.v. page 308). Table 6.6 lists the data for the two groups, Control and Treatment, with $n = 23$ in each group. Again, the $n = 23$ data points in the Control group are fixed, and one value, w, in the Treatment group is allowed to vary. Table 6.7 lists eleven values for w ranging from a low value of $w = 97$ up to a high value of $w = 497$, the exact two-sided probability values for the Fisher–Pitman matched-pairs permutation test with $v = 1$ and $v = 2$, the exact two-sided probability values for the permutation version of the Wilcoxon matched-pairs rank-sum test [1453],[32] and the two-sided probability values for the classical matched-pairs t test. Each of the exact two-sided probability values in Table 6.7 is based on

$$M = 2^n = 2^{23} = 8{,}388{,}608$$

[32] The exact Fisher–Pitman matched-pairs permutation tests with $v = 1, 2$ and the exact permutation version of the Wilcoxon matched-pairs rank-sum test are specific forms of MRBP with $g = 2$ and $r = 1$ (qq.v. pages 310 and 317).

Table 6.6 Raw-score observed values for matched-pairs control and treatment groups with $n = 23$ objects in each group

Control	Treatment	Control	Treatment	Control	Treatment
287	288	269	275	290	293
270	273	294	298	281	285
287	291	273	276	274	276
283	284	289	286	294	295
271	269	291	296	267	271
291	294	267	268	283	278
280	281	285	287	294	w
290	292	267	272		

Table 6.7 Two-sided probability value comparisons for the exact Fisher–Pitman matched-pairs test with $v = 1$ and $v = 2$, the exact Wilcoxon matched-pairs rank-sum test, and the classical matched-pairs t test

	Permutation test		Wilcoxon	Classical
w	$v = 1$	$v = 2$	rank test	t test
97	0.00636	0.99909	0.01946	0.45984
197	0.00636	0.99909	0.01946	0.60982
247	0.00636	0.99961	0.01946	0.96883
272	0.00636	0.50154	0.01946	0.41245
287	0.00083	0.02746	0.01946	0.04000
297	0.00077	0.00183	0.00193	0.00707
307	0.00077	0.00154	0.00164	0.01038
322	0.00077	0.00154	0.00164	0.03623
347	0.00077	0.00154	0.00164	0.09660
397	0.00077	0.00154	0.00164	0.17852
497	0.00077	0.00154	0.00164	0.24619

equally-likely arrangements of the data given in Table 6.6. The two-sided probability values for the classical matched-pairs t test are based on Student's t distribution with $n - 1 = 23 - 1 = 22$ degrees of freedom, under the usual assumptions of normality and independence.

As with the independent two-sample tests, the exact two-sided probability values for the Fisher–Pitman matched-pairs permutation test with $v = 1$ are stable, consistent, and are unaffected by the extreme values of w in either direction. Similar results are achieved with the Wilcoxon matched-pairs rank-sum test. In contrast, the exact two-sided probability values for the Fisher–Pitman matched-pairs permutation test with $v = 2$ range from 0.99909 for small values of w down to 0.00154 for large values of w, relative to the fixed values. The two-sided probability values for the classical matched-pairs t test are inconsistent, with low probability values associated with moderate values of w and higher probability values with both low and high values of w. Again, with extreme values, the classical matched-pairs t test is unable to detect the obvious differences in location between the matched Control and Treatment groups. In conclusion, Mielke et al. explained that the examples clearly showed the advantages of (1) utilizing a Euclidean distance function, (2) employing a permutation approach for data analysis [978, p. 212], and (3) avoiding the loss of information associated with substituting rank-order statistics for numerical measurements.

6.17 Advantages of the Median for Analyzing Data

The median possesses certain advantages relative to the arithmetic, geometric, and harmonic means for both describing and analyzing data. Let $x_1, \ldots, x_n$ denote n observed values in an ordinary Euclidean data space, where n is not exceedingly small. Let S_1, S_2, S_3, and S_4 denote the median, arithmetic mean, geometric mean, and harmonic mean, respectively, and let w denote a specified value that is arbitrarily changed while the remaining $n - 1$ values remain fixed, as in Tables 6.4 and 6.6. Finally, assume that (1) the interval of the observed values comprising S_1 and S_2 is $-\infty < x_i < +\infty$ for $i = 1, \ldots, n$, and (2) the interval of the observed values comprising S_3 and S_4 is $0 < x_i < +\infty$ for $i = 1, \ldots, n$.

Specific formulae for S_1, S_2, S_3, and S_4 are:

$$S_1 = M = \text{median}\,(x_1, \ldots, x_n)\,,$$

$$S_2 = \frac{1}{n} \sum_{i=1}^{n} x_i\,,$$

$$S_3 = \left(\prod_{i=1}^{n} x_i \right)^{1/n}\,,$$

and

$$S_4 = n \left(\sum_{i=1}^{n} x_i^{-1} \right)^{-1}\,.$$

Then,

$$\lim_{w \to +\infty} S_1 = \lim_{w \to -\infty} S_1 = M,\ \lim_{w \to +\infty} S_2 = +\infty,\ \lim_{w \to -\infty} S_2 = -\infty,$$

$$\lim_{w \to 0} S_3 = 0,\ \lim_{w \to +\infty} S_3 = +\infty,\ \text{and}\ \lim_{w \to 0} S_4 = 0,\ \lim_{w \to +\infty} S_4 = V,$$

where V is the resulting value of S_4 when $w^{-1} = 0$. Consequently, S_1 is not influenced by a few isolated extreme values, whereas S_2, S_3, and S_4 can be substantially influenced by even a few extreme values.

In addition, the statistical inferences associated with S_1, S_2, S_3, and S_4 behave in a similar manner. The data space of observed values $x_1, \ldots, x_n$ is again an ordinary

Euclidean space that satisfies the three properties of a metric space. If $\Delta(x_i, x_j)$ denotes the distance function between x_i and x_j, then:

1. $\Delta(x_i, x_j) \geq 0$ and $\Delta(x_i, x_i) = 0$.
2. $\Delta(x_i, x_j) = \Delta(x_j, x_i)$, termed "symmetry."
3. $\Delta(x_i, x_j) + \Delta(x_j, x_k) \geq \Delta(x_i, x_k)$, termed "the triangle inequality."

The analysis space of S_1 is also an ordinary Euclidean metric space with the absolute distance function given by $\Delta_1(x_i, x_j) = |x_i - x_j|$ and is congruent with the Euclidean-data metric space consisting of $x_1, \ldots, x_n$ [938, pp. 815 and 820].

However, the analysis space of S_2 is a squared Euclidean space with the squared Euclidean distance function given by $\Delta_2(x_i, x_j) = (x_i - x_j)^2$ and is not congruent with the ordinary Euclidean space since property (3) of a metric space is not satisfied. For example, if $x_1 = 4$, $x_2 = 5$, and $x_3 = 6$ of the squared Euclidean space, then $\Delta_2(x_1, x_2) = \Delta_2(x_2, x_3) = 1$ and $\Delta_2(x_1, x_3) = 4$ demonstrates that the triangle inequality of a metric space is not satisfied. Similarly, the distance functions of S_3 and S_4 are $\Delta_3(x_i, x_j) = (\ln x_i - \ln x_j)^2$ and $\Delta_4(x_i, x_j) = (x_i^{-1} - x_j^{-1})^2$, respectively, are also associated with non-metric distance functions. If a value of either x_i or x_j approaches 0, then both $\Delta_3(x_i, x_j)$ and $\Delta_4(x_i, x_j)$ approach ∞. Thus, only S_1 has an analysis space that is congruent with the ordinary Euclidean metric space of the observed values. Results of statistical inference procedures involving data and analysis spaces that differ are naturally questionable regarding any interpretation.

For these reasons, S_1 appears to be a more natural choice in describing and analyzing data than S_2, S_3, or S_4. A disconcerting fact pertaining to this discussion is that the most commonly-used statistical inference methods, such as the t and F tests, are based on an ordinary Euclidean metric data space and a non-metric analysis space associated with S_2 and, consequently, are not congruent. When the geometries of the data and analysis spaces of a statistical method are not congruent, the results of such a statistical method are questionable since there is a lack of correspondence between the data and analysis spaces [938, 965].

6.18 Consideration of Statistical Outliers

The subject of how to treat statistical outliers, extreme values, or as John Tukey referred to them, "wild observations," has been discussed in the statistical literature for well over 100 years; see for example, two papers in 1960 by Anscombe [27] and Daniel [312]. The problem is still serious as Higgins and Blair noted in 2000:

> [w]hen observations are from heavier-tailed distributions, a few extreme observations can diminish the power of a means-based statistic to detect differences between treatments. The problem exists whether one assumes the population-sample model or the randomization model [614, p. 86],

and Yadolah Dodge, writing on L_1 estimators in 1987, wrote:

> [w]hile the method of least squares (and its generalizations) have served statisticians well for a good many years (mainly because of mathematical convenience and ease of

> computation), and enjoys certain well known properties within strictly Gaussian parametric models, it is recognized . . . that outliers, which arise from heavy-tailed distributions, have an unusually large influence on estimates obtained by these methods. Indeed, one single outlier can have an arbitrary [sic] large effect on the estimate [354, p. 3].

Of course, the subject of outliers has an even longer history in the astronomy literature. Writing in 1777, the eighteenth century mathematician and physician, Daniel Bernoulli directed this comment at astronomers:

> [n]evertheless, I do not condemn in every case the principle of rejecting one or other of the observations, indeed I approve it, whenever in the course of observation an accident occurs which in itself raises an immediate scruple in the mind of the observer, before he has considered the event and compared it with the other observations. If there is no such reason for dissatisfaction I think each and every observation should be admitted whatever its quality, as long as the observer is conscious that he has taken every care [105, 735] (Bernoulli, quoted in Finney [438, p. 311]).

An Illustration

To illustrate the deletion of discordant values in astronomy, consider the work of James Short [1268, 1269]; see also books on this topic by Lomb [839] and Sheehan and Westfall [1257]. At about the same time that Daniel Bernoulli was writing, James Short, Fellow of the Royal Society and maker of reflecting telescopes, observed the 1761 transit of Venus from Savile House in Leicester Square, London. He was there at the invitation of his Royal Highness, the Duke of York, who was present for the transit with a number of members of the royalty: his Royal Highnesses Prince William, Prince Henry, Prince Frederick, and her Royal Highness Lady Augusta [1268, p. 180].

Short began his analysis with the timings of Venus leaving the disk of the Sun (internal contact at egress) from a number of different locations, comparing the timings of Mason and Dixon at the Cape of Good Hope with that of 15 timings from Europe. His final value for the Sun–Earth distance or parallax was the mean of these, leaving out a few outliers. Based on the 15 observations, he found the mean value to be $8''$, 47. After deleting the four observations that differed the most from the mean, he found a corrected mean value to be $8''$, 52, which translates to 152.1 million km (94.5 million miles), a value not far removed from the present-day value of approximately 149.6 million km (92.9 million miles) [839, p. 75]. The four observations Short deleted were from Shirburn Castle, Oxfordshire ($8''$, 15); Tornea, Finland ($8''$, 07); Drontheim, Norway ($8''$, 23); and Calmar (Kalmar), Sweden ($8''$, 86) [1269, p. 615]. James Short F.R.S. died in Newington Butts, London, on 15 June 1768 at the age of 58.

An outlier, such as observation value w in Tables 6.4 and 6.6, may be defined as a value which seems either too large or too small as compared to the rest of the observed values [563, p. 165].[33,34] Yadolah Dodge, in *The Oxford Dictionary of Statistical Terms* provides a more comprehensive definition:

> [i]n a sample of n observations, it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is at fault. Such values are called outliers. Tests are available to ascertain whether, under certain assumptions, they can be accepted as homogeneous with the rest of the sample [356, p. 297] (also quoted in Finney [438, p. 310]).

In 2002 Roy investigated the effects of heteroscedasticity and outliers on the size and power of the permutation t test for small-sample problems [1200]. Roy used sample sizes of 10, 25, and 50 and Cauchy, folded normal, log-normal, and normal distributions with outliers in a computer simulation based on 10,000 Monte Carlo samples. In general, Roy concluded that unless there is very high kurtosis, an appreciable proportion of outliers, or very small samples, the permutation t test performed very well in terms of size and power, even when heteroscedasticity was present [1200, p. 26].

In general, the exclusion of outliers on a purely statistical basis has been, and remains, a dangerous procedure [563]. Both John Tukey [1378] and William Kruskal [777] suggested using mixtures of normal distributions with the same mean but different variances, while myriad others have advocated non-parametric approaches; Philip Good, for example, recommended the use of ranks when outliers are a concern [529, 530]. Tukey also considered truncation and Winsorizing as potential solutions [1378]. The problem exists in part, of course, because of the use of mean-based statistical procedures in which outliers, and other values, are weighted by the squares of their deviations from the mean, i.e., $v = 2$, thereby increasing their influence proportional to their squared deviations from the mean; see also a 1981 article by Bert Green in *Journal of the American Statistical Association* [549] (q.v. page 337). A median-based statistical procedure with the substitution of absolute deviations from the median, i.e., $v = 1$, mitigates the problem as extreme values are no longer weighted by their squares.[35]

While the two example analyses in Tables 6.5 and 6.7, based on the data listed in Tables 6.4 and 6.6, demonstrate that the Euclidean distance ($v = 1$) function tests eliminate the need for replacing the observed raw data in question with rank-order statistics to accommodate the existence of extreme events when $v = 2$, the two examples also address an older question when squared Euclidean distance ($v = 2$)

[33] In 2008 Malcolm Gladwell published an entire book titled *Outliers*, in which he defines an outlier as "a statistical observation that is markedly different in value from the others of the sample" [515, p. 3].

[34] Two lucid discussions of outliers and how to treat them are contained in papers by David Finney in 2006 [438] and John Ludbrook in 2008 [853].

[35] Kruskal suggested that only identified outliers be given a lesser weight, then proceeding as usual [777]. Others who earlier suggested the weighting of outliers include S. Newcomb [1032] who suggested that each observation be weighted by its residual, E.G. Stone [1325], and F.Y. Edgeworth [380, 393].

functions are used. The Fisher–Pitman permutation versions of the t tests may differ substantially from the classical t tests based on normality in contrast with references suggesting otherwise; see for example, an article by Boik in 1987 [175]. Since the observed responses are perceived in a Euclidean data space, it is natural to have the analysis space congruent with the Euclidean data space, i.e., $v = 1$ [938, 939, 941].

6.19 Multivariate Multiple Regression Analysis

In 2002 and 2003, extensions of multiple regression permutation analyses to applications involving multivariate dependent values were considered by Mielke and Berry [963, 964]. The extensions were prompted by a multivariate Least Sum of Euclidean Distances (LSED) algorithm developed by Kaufman, Taylor, Mielke, and Berry in 2002 [711]. Consider the multiple regression model given by

$$y_{Ik} = \sum_{j=1}^{m} x_{Ij}\beta_{jk} + e_{Ik} \tag{6.1}$$

for $I = 1, \ldots, N$ and $k = 1, \ldots, r$, where y_{Ik} denotes the Ith of N measurements for the kth of r response variables, possibly affected by a treatment; x_{Ik} is the jth of m covariates associated with the Ith response, where $x_{Ij} = 1$ indicates that the model includes an intercept; β_{jk} denotes the jth of m regression parameters for the kth of r response variables; and e_{Ik} designates the error associated with the Ith of N measurements for the kth of r response variables. If the estimates of β_{jk} that minimize

$$\sum_{I=1}^{N} \left(\sum_{k=1}^{r} e_{Ik}^2 \right)^{1/2}$$

are denoted by $\tilde{\beta}_{jk}$ for $j = 1, \ldots, m$ and $k = 1, \ldots, r$, then the N r-dimensional residuals of the multivariate multiple regression model based on LSED are given by

$$\tilde{e}_{Ik} = y_{Ik} - \sum_{j=1}^{m} x_{Ij}\tilde{\beta}_{jk}$$

for $I = 1, \ldots, N$ and $k = 1, \ldots, r$. In comparison to multivariate multiple regression models that minimize

$$\sum_{I=1}^{N}\sum_{k=1}^{r} |e_{Ik}| , \qquad \sum_{I=1}^{N}\left(\sum_{k=1}^{r} |e_{Ik}| \right)^2 , \qquad \text{or} \qquad \sum_{I=1}^{N}\sum_{k=1}^{r} e_{Ik}^2 ,$$

only the multivariate multiple regression model based on LSED does not vary with coordinate rotation and possesses the desired geometrical attributes of satisfying the triangle inequality of a metric [711, 942].

6.19.1 A Permutation Test

Mielke and Berry utilized multi-response permutation procedures (MRPP) to analyze the residuals from a multivariate multiple regression analysis (q.v. page 254). Let the N r-dimensional residuals, $\tilde{e}_{I1}, \ldots, \tilde{e}_{Ir}$, for $I = 1, \ldots, N$ obtained from a multivariate multiple regression model based on LSED be partitioned into g treatment groups of sizes $n_1, \ldots, n_g$, where $n_i \geq 2$ for $i = 1, \ldots, g$ and $\sum_{i=1}^{g} n_i = N$. The MRPP analysis of the multivariate residuals depends on the statistic

$$\delta = \sum_{i=1}^{g} C_i \xi_i ,$$

where $C_i = n_i/N$ is a positive weight for the ith of g treatment groups that minimizes the variability of δ, $\sum_{i=1}^{g} C_i = 1$, and ξ_i, the average pairwise Euclidean distance among the n_i r-dimensional residuals in the ith of g treatment groups, is defined by

$$\xi_i = \binom{n_i}{2}^{-1} \sum_{K=1}^{N-1} \sum_{L=K+1}^{N} \left[\sum_{j=1}^{r} \left(\tilde{e}_{Kj} - \tilde{e}_{Lj} \right)^2 \right]^{v/2} \Psi_{Ki} \Psi_{Li} ,$$

where $v > 0$,

$$\Psi_{Ii} = \begin{cases} 1 & \text{if } \tilde{e}_{I1}, \ldots, \tilde{e}_{Ir} \text{ is in the } i\text{th of } g \text{ treatment groups,} \\ 0 & \text{otherwise,} \end{cases}$$

and $v = 1$ yields Euclidean distance. The null hypothesis specifies that each of the

$$M = \frac{N!}{\displaystyle\prod_{i=1}^{g} n_i!}$$

possible allocations of the N r-dimensional residuals to the g treatment groups is equally likely. Under the null hypothesis, the permutation distribution of δ assigns equal probabilities to the resulting M values of δ. Since small values of δ imply a concentration of similar residuals within the g treatment groups, the null hypothesis is rejected when the observed value of δ, δ_{o}, is small. Thus, the exact MRPP probability (P) value associated with δ_{o} is given by

$$P\left(\delta \leq \delta_{\text{o}} | H_0\right) = \frac{\text{number of } \delta \text{ values} \leq \delta_{\text{o}}}{M} .$$

In addition, approximate MRPP probability values may be obtained from either resampling-approximation or Pearson type III moment-approximation algorithms (q.v. page 303). Compared with classical parametric approaches, the Euclidean-distance multivariate multiple regression method is exceedingly robust with regard to extreme values and, being a permutation test, does not depend on assumptions such as normality, homogeneity, and independence. Further, the permutation approach allows for the choice of exact or approximate probability values. Other applications of the multivariate multiple regression method include various completely randomized and randomized block designs such as one-way, Latin squares, factorial, nested, and split-plot designs, both with and without covariates [965, Chap. 5]. Unlike parametric procedures, the only required assumption is the random assignment of treatments to subjects.

Univariate multiple regression analyses based on LSED and MRPP were originally introduced in 1982 using rank-order transformations of the observed raw residuals and in 1983 using the observed raw residuals, the preferred choice due to the robustness of MRPP based on a Euclidean distance function [981, 1468]. More specifically, Mielke, Berry, and Medina analyzed wintertime orographic cloud seeding experiments high in the Colorado mountains (Climax I and II), publishing the results in *Journal of Applied Meteorology* [981]. The authors identified two problems associated with the use of the classical linear model fitted by least squares to analyze the data: (1) the residual data to be analyzed were highly skewed and heavily dependent on only a few very large values, and (2) the complex non-Euclidean geometry underlying the classical linear model. Consequently, Mielke et al. utilized an alternative analysis procedure whereby they produced residual data from a median (least absolute deviation or LAD) regression model and a subsequent analysis based on rank tests associated with MRPP utilizing a Euclidean distance function (i.e., $v = 1$) [981]. At the time the authors were unaware of the robustness of MRPP with $v = 1$ and the artificial nature of rank-order statistics.

The following year, Wong, Chidambaram, and Mielke executed the identical regression analysis on surface hail observations taken in 1975 and 1976 in Alberta, Canada, publishing the results in *Atmosphere–Ocean* in 1983 [1468]. However, in this case they utilized the observed (raw) residuals and did not convert the residuals to ranks. Thus, Wong et al. were the first to present modern regression analyses based on median (LAD) regression that was combined with MRPP applied to the observed residual data and utilizing a Euclidean distance function ($v = 1$) [1468].

In 2005 Endler and Mielke utilized multivariate multiple regression to compare entire color patterns as birds actually see them [410]. Noting that color patterns and their visual backgrounds consist of a mosaic of patches that vary in color, brightness, size, shape, and position, they used the LSED–MRPP method to compare entire color patterns instead of comparing multiple pairs of patches as was customary in previous studies. They observed that the LSED–MRPP method has two desirable features: (1) it satisfies the congruence principle, i.e., that the metric Euclidean distance analysis space is congruent with the observed metric Euclidean distance data space, and (2) the Euclidean distance predicts perceived color distances [410, p. 418]. They explained, this is in contrast to the classical t and F tests that are based on non-metric squared Euclidean distance analysis spaces and do not satisfy

Table 6.8 Bivariate data on Scholastic Competence and Global Self-worth (y_1, y_2) for an unbalanced randomized block design with $N = 16$ students

	University		
Academic year	A	B	C
Freshman	155, 144		170, 128
	156, 139		167, 131
	187, 100		173, 121
	152, 147		176, 121
	161, 142		
Senior	162, 133	177, 119	175, 122
	157, 136	173, 123	
		184, 115	
		180, 118	

the congruence principle, thus bearing no simple relationship to expected perceptual differences.

6.19.2 An Example Analysis

To illustrate a residual permutation analysis, consider an unbalanced randomized block experimental design, where scores were collected within three universities (A, B, C) at two time periods (Freshman and Senior years) on two scales of a standardized test (Scholastic Competence and Global Self-worth). The data are summarized in Table 6.8 for a small sample of $N = 16$ students. Although the residual permutation analysis can easily accommodate many dimensions, larger numbers of subjects, and more complicated designs, the example is intentionally kept simple to illustrate the procedures.

6.19.2.1 Analysis of Universities

The model under the null hypothesis for the analysis of universities is given by

$$y_{Ik} = x_{I1}\beta_{1k} + x_{I2}\beta_{2k} + e_{Ik},$$

where $I = 1, \ldots, 16$ and $k = 1, 2$ correspond to Eq. (6.1). The values for the $\mathbf{X}$ (dummy coded) and $\mathbf{Y}$ (data) matrices (x_{Ij} and y_{Ik}) are given in Table 6.9 where $x_{I1} = 1$ for the intercept, $x_{I2} = 1$ (0) for the Freshman (Senior) academic years, and y_{Ik} denotes the kth response of the Ith student. If δ_{o} denotes the observed value of δ, then the exact permutation analysis based on $M = 1{,}441{,}440$ permutations yields $\delta_{\text{o}} = 12.8587$ with an exact probability value of $6{,}676/1{,}441{,}440 \doteq 0.4631 \times 10^{-2}$, a resampling-approximation permutation procedure based on $L = 1{,}000{,}000$ yields $\delta_{\text{o}} = 12.8587$ with an approximate resampling probability value of 0.4689×10^{-2}, and a Pearson type III moment-approximation permutation analysis (q.v. page 303) yields $\delta_{\text{o}} = 12.8587$ with an approximate probability value of 0.4701×10^{-2}.

Table 6.9 Data file containing the **X** and **Y** matrices for the analysis of universities

University	**X** matrix		**Y** matrix	
A	1	1	155	144
	1	1	156	139
	1	1	187	100
	1	1	152	147
	1	1	161	142
	1	0	162	133
	1	0	157	136
B	1	0	177	119
	1	0	173	123
	1	0	184	115
	1	0	180	118
C	1	1	170	128
	1	1	167	131
	1	1	173	121
	1	1	176	121
	1	0	175	122

6.19.2.2 Analysis of Academic Years

The model under the null hypothesis for the analysis of academic years is given by

$$y_{Ik} = x_{I1}\beta_{1k} + x_{I2}\beta_{2k} + x_{I3}\beta_{3k} + e_{Ik},$$

where $I = 1, \ldots, 16$ and $k = 1, 2$ correspond to Eq. (6.1). The values of the **X** (dummy coded) and **Y** (data) matrices (x_{Ij} and y_{Ik}) are given in Table 6.10 where $x_{I1} = 1$ for the intercept, (x_{I2}, x_{I3}) is (1, 0), (0, 1), and (0, 0) for University A, B, and C, respectively, and y_{Ik} denotes the kth response of the Ith student. The exact permutation analysis based on $M = 11{,}440$ permutations yields $\delta_o = 12.0535$ with an exact probability value of $2{,}090/11{,}440 \doteq 0.1827$, a resampling-approximation permutation procedure based on $L = 1{,}000{,}000$ yields $\delta_o = 12.0535$ with an approximate resampling probability value of 0.1826, and a Pearson type III moment-approximation permutation analysis (q.v. page 303) yields $\delta_o = 12.0535$ with an approximate probability value of 0.1901.

6.20 O'Gorman and Multiple Linear Regression

In 2005 O'Gorman evaluated the performance of randomization tests that use permutations of independent variables in multiple linear regression models [1050]. In this paper O'Gorman introduced a new permutation method that he called the permute-Z method. A little background is in order.

There has long existed a controversy over the appropriate permutation method for analyzing multiple linear regression models. Consider the multiple linear regression model given by

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X} + \beta_2\mathbf{Z} + \boldsymbol{\varepsilon}, \tag{6.2}$$

Table 6.10 Data file containing the **X** and **Y** matrices for the analysis of academic years

Academic year	X matrix			Y matrix	
Freshman	1	1	0	155	144
	1	1	0	156	139
	1	1	0	187	100
	1	1	0	152	147
	1	1	0	161	142
	1	0	0	170	128
	1	0	0	167	131
	1	0	0	173	121
	1	0	0	176	121
Senior	1	1	0	162	133
	1	1	0	157	136
	1	0	1	177	119
	1	0	1	173	123
	1	0	1	184	115
	1	0	1	180	118
	1	0	0	175	122

where $\mathbf{Y}$ is an $n \times 1$ vector of dependent variables, $\mathbf{X}$ and $\mathbf{Z}$ are $n \times 1$ vectors of independent variables, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors. Also define a reduced regression model given by

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \boldsymbol{\varepsilon}' . \tag{6.3}$$

In 1983 Freedman and Lane proposed permuting the residuals obtained from the reduced regression model in Eq. (6.3) [478]. The permuted residuals were then added to the predicted values calculated from the reduced regression model to form new dependent values, which were then subjected to the full regression model in Eq. (6.2) to obtain the t statistic for testing $H_0\colon \beta_2 = 0$. Eleven years later in 1992, ter Braak proposed an alternative permutation procedure that was very similar to that of Freedman and Lane, except that ter Braak computed the residuals from the full regression model in Eq. (6.2) [1346].

Two years later in 1995, Kennedy proposed another method of permutation that he claimed was identical to the permutation method of Freedman and Lane [748]. The Kennedy method correlated the residuals of the regression of **Y** and **X** with the residuals of the regression of **Z** and **X**. The test is based on the idea that the partial regression coefficient is equivalent to the simple regression coefficient of residuals [21, p. 78].

In 1997 Manly proposed simply permuting the observed values of **Y** for the test of partial correlation [876]. In 1999 Anderson and Legendre evaluated the four permutation methods of Freedman and Lane, ter Braak, Kennedy, and Manly, concluding that the reduced regression model of Freedman and Lane came closer to maintaining the level of significance than the other three methods [20]. In 2001 Anderson and Robinson described the asymptotic properties of the four methods [21].

The permute-Z method advocated by O'Gorman was not new. As pointed out by O'Gorman, the permute-Z method was first used to test $H_0 \colon \beta_2 = 0$ by Draper and Stoneman in 1966 [360]. Also, O'Gorman noted that Kennedy and Cade had used the permute-Z method in a small simulation study to demonstrate that the type I error of the permute-Z method approximates the nominal value if a t statistic is used as the permutation statistic [749]. Kennedy and Cade called it the "shuffle-Z" method (q.v. page 351).

The permute-Z method advocated by O'Gorman is easy to describe and consists of four parts [1050, p. 898]:

1. For the raw data, compute a conventional F test statistic for a subset of regression coefficients from the regression of Y on X, and denote the result by F^*.
2. For each permutation, permute the rows of Z and use the full model to obtain the test statistic.
3. Generate R permutations of the rows of Z, and let E be the number of times that the permutation test statistic exceeds F^*.
4. Compute the probability value as $p = (E + 1)/(R + 1)$ and reject H_0 if $p \leq \alpha$.

O'Gorman evaluated the performance of the four methods using an extensive simulation study. He showed that the permute-Z method maintained its level of significance, except for extreme situations, and had power that approximated the power of the reduced-model test proposed by Freedman and Lane. Furthermore, he showed, by way of an example, that the permute-Z method can be more valuable than the Freedman–Lane test in its ability to "downweight" outliers [1050].

6.21 Brusco–Stahl–Steinley and Weighted Kappa

In 2008 Brusco, Stahl, and Steinley presented an implicit enumeration method for an exact permutation test of Cohen's weighted kappa measure [226] (q.v. page 382). Noting that complete enumeration of all possible agreement tables, given fixed marginal frequency totals, is computationally unwieldy for modest numbers of objects and categories, they proposed an implicit enumeration algorithm for conducting an exact permutation test of Cohen's weighted kappa, which was applicable to agreement tables of non-trivial size.

The problem, they explained, is that when using resampling-approximation permutation methods, the number of samples necessary to obtain a good probability approximation must be quite large whenever the actual probability is small, which often occurs when the computed value of the observed weighted kappa is high, e.g., $\kappa \geq 0.50$. The suggested procedure was to examine partially filled sampled tables and to "prune" those tables that could not produce a weighted kappa value greater than the weighted kappa value of the observed table. The process is to place cell frequencies into those cells with the smallest weights first, thereby quickly obtaining partially constructed tables that cannot possibly achieve the value of the observed weighted kappa. This appears at first reading to be similar to the network algorithm of Mehta and Patel (q.v. page 287). However, the algorithm of Brusco et al. is not a

Fig. 6.6 Multiple sclerosis rater agreement panel for $N = 149$ patients in Winnipeg, Manitoba

		Winnipeg				
		1	2	3	4	Total
New Orleans	1	38	5	0	1	24
	2	33	11	3	0	47
	3	10	14	5	6	35
	4	3	7	3	10	23
	Total	84	37	11	17	149

network algorithm and is based on $(k-1)^2$ nested loops, where k is the number of ordered categories for classifying n objects by each of the two judges.

An example will illustrate the process. Following Brusco et al., considered the agreement table in Fig. 6.6, which was originally published by Landis and Koch in *Biometrics* in 1977 [796] and is based on data originally collected by Westlund and Kurland and published in *American Journal of Hygiene* in 1953 [1439].[36] Two neurologists, one in Winnipeg, Manitoba, and one in New Orleans, Louisiana, reviewed the records of $n = 149$ patients in Winnipeg and independently placed them into one of $k = 4$ ordered diagnostic categories: (1) certain multiple sclerosis, (2) probable multiple sclerosis, (3) possible multiple sclerosis, and (4) (doubtful/unlikely/definitely not) multiple sclerosis. Based on a quadratic weighting scheme, where w_{ij} denotes the weights given by

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \tag{6.4}$$

for $i, j = 1, \ldots, k$, they calculated the variable portion of the standard kappa formula,

$$[\mathbf{WT}] = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij} ,$$

where x_{ij} indicates an observed cell frequency, $i, j = 1, \ldots, k$.

Now consider the first random table generated with the same marginals, as depicted in Fig. 6.7, in which just two cells have been filled, i.e., cells $\{1, 4\}$ and $\{4, 1\}$, where $x_{1,4} = 8$ and $x_{4,1} = 11$. The fundamental insight for the implicit enumeration scheme stems from the fact that no completion of the remaining cells in the rater agreement table in Fig. 6.7 can possibly produce a weighted kappa statistic that equals or exceeds the observed statistic with $[\mathbf{WT}] = 130.33$ and $\kappa = 0.5246$ [226, p. 444]. As noted by Brusco et al., although a large number of completed tables satisfying the marginals can be realized by filling in the remaining cells of the agreement table in Fig. 6.7, none of these will contribute to the probability value.

[36] Others have analyzed the Westlund and Kurland data, including Jolayemi in 1990 [698] and Borkowf in 2004 [183].

Fig. 6.7 Partially constructed rater agreement table for the multiple sclerosis data

		Winnipeg				
		1	2	3	4	Total
	1	—	—	—	8	24
New	2	—	—	—	—	47
Orleans	3	—	—	—	—	35
	4	11	—	—	—	23
Total		84	37	11	17	149

Therefore, the partially constructed rater-agreement table, and all complete tables stemming from the partial table, can be eliminated from further consideration. In this manner, Brusco et al. demonstrated that their enumeration algorithm provided tremendous computational savings when the number of objects rated is relatively small ($n \leq 150$) and the observed weighted kappa statistic is at least moderately high ($\kappa \geq 0.50$).[37]

The efficiency of the implicit enumeration algorithm of Brusco et al. stems from the ordering of nested loops and the pruning rule. Application of the algorithm to the data in Fig. 6.6 yielded the same exact probability value of 1.3467×10^{-14} that was found by complete enumeration, but took only about 2 % of the computing time. The reason for the computational savings was that the implicit enumeration algorithm required evaluation of only 46,980,632 (1.49 %) of the possible 3,146,622,222 tables with the same marginal frequency totals as the observed $k \times k$ contingency table [226, p. 445].

6.22 Mielke et al. and Ridit Analysis

The classical two-treatment ridit analysis was first introduced by I.D.J. Bross in 1958 [223]. Ridit is an acronym for *R*elative to an *I*dentified *D*istribution, where the suffix "it" represents a type of data transformation similar to logit and probit. The most common application of ridit analysis compares two independent treatment groups in which ridit scores are calculated for the c ordered category frequencies of the first treatment group and applied to the c ordered categories of the second treatment group, and vice-versa. In this application, the two treatment groups are considered to be independent finite samples.

In 2009 Mielke, Long, Berry, and Johnston extended the two-treatment ridit analysis developed by Bross to $g \geq 2$ treatment groups [986]. Following Mielke et al., consider a $c \times g$ cross-classification table with c ordered disjoint response categories and g unordered treatment categories. Let m_{ij} denote the observed cell frequency of the ith row and the jth column for $i = 1, \ldots, c$ and $j = 1, \ldots, g$.

[37] Note that Brusco et al., in order to ensure an efficient procedure, began the process by filling in those cells with the smallest weights; in this case, following Eq. (6.4), $w_{ij} = w_{1,4} = 1 - (1 - 4)^2/(4-1)^2 = w_{4,1} = 1 - (4-1)^2/(4-1)^2 = 0$.

Also, let

$$M_j = \sum_{i=1}^{c} m_{ij}$$

denote the unordered treatment frequency totals for $j = 1, \ldots, g$, let

$$L_i = \sum_{j=1}^{g} m_{ij}$$

denote the ordered response frequency totals for $i = 1, \ldots, c$, and let

$$N = \sum_{i=1}^{c} \sum_{j=1}^{g} m_{ij}$$

denote the table frequency total for the cg cells. The ridit scores of the jth observed treatment group, $j = 1, \ldots, g$, are then given by

$$\begin{aligned}
R_{1j} &= m_{1j}/(2M_j), \\
R_{2j} &= (m_{1j} + m_{2j}/2)/M_j, \\
&\vdots \qquad\qquad \vdots \\
R_{cj} &= (m_{1j} + \cdots + m_{c-1,j} + m_{cj}/2)/M_j .
\end{aligned}$$

Finally, define a ridit test statistic, T, based on differences among all possible pairs of treatment groups given by

$$T = \sum_{i=1}^{g-1} \sum_{j=i+1}^{g} \left| x_{ij} - x_{ji} \right| ,$$

where

$$x_{ij} = \sum_{k=1}^{c} \frac{R_{ki} m_{kj}}{M_j}$$

for $i, j = 1, \ldots, g$.

As noted by Mielke et al., ridit scores possess a probabilistic interpretation, i.e., the ridit score R_{ij} for the ith of c ordered categories in the jth treatment group is the proportion of observations in the categories below the ith category in the jth treatment group, plus half the proportion of observations in the ith category of the jth of g treatment groups. Thus, R_{ij} is the probability that a randomly selected observation falls below the midpoint of the ith category, conditioned on the jth treatment [986, p. 225]. Figure 6.8 illustrates the calculation of ridit scores for one

Fig. 6.8 Example calculation of ridit scores from $c = 5$ ordered categories

Category	Frequency	Ridit
Strongly agree	2	0.0263
Agree	16	0.2632
Neutral	7	0.5658
Disagree	5	0.7237
Strongly disagree	8	0.8947
Total	38	

treatment group and $c = 5$ ordered categories. In Fig. 6.8 , $m_{1j} = 2/(2)(38) = 0.0263$, $m_{2j} = (2 + 16/2)/38 = 0.2632$, $m_{3j} = (2 + 16 + 7/2)/38 = 0.5658$, and so on.

An exact permutation test is usually impractical for a ridit analysis, as among the c^N equally-likely assignment configurations under the null hypothesis that the g treatments come from a common population and all possible outcomes of the ridit analysis are equally likely, there are

$$W = \prod_{j=1}^{g} \binom{M_j + c - 1}{c - 1}$$

distinguishable partitions of the c^N configurations of the g treatment groups. Thus, for $c = 5$, $g = 4$, and $M_1 = M_2 = M_3 = M_4 = 15$,

$$W = \binom{15 + 5 - 1}{5 - 1}^4 = (3{,}876)^4 = 225{,}701{,}826{,}437{,}376$$

and $c^N = 5^{60} \doteq 8.67 \times 10^{41}$, which is a very large number.[38]

A resampling-approximation permutation procedure generates L sets of N random assignments selected from the c^N assignment configurations of the g treatment groups. A ridit test statistic T is then calculated for each of the L sets of N random assignments of the ordered category frequencies. Given the resampled ridit statistics $T_1, \ldots, T_L$, the resampling-approximation upper-tail probability value, P, of the observed value of T, T_{o}, under the null hypothesis is given by

$$P = \frac{1}{L} \sum_{i=1}^{L} \Phi(T_i),$$

[38] Actually, $5^{60} = 867{,}361{,}737{,}988{,}403{,}547{,}205{,}962{,}240{,}695{,}953{,}369{,}140{,}625$.

where

$$\Phi(T_i) = \begin{cases} 1 & \text{if } T_i \geq T_\text{o}, \\ 0 & \text{otherwise.} \end{cases}$$

Mielke et al. concluded the article with an example analysis based on $g = 4$ opoids (Fentanyl, Codeine, Oxycodone, and Morphine), classified into $c = 5$ ordered evaluation categories (Excellent, Good, Adequate, Weak, and Poor), and administered to $N = 149$ patients, each of whom had received a robotic-assisted laparoscopic radical prostatectomy and was randomly assigned to one of the four post-surgery treatment groups. Based on $L = 1{,}000{,}000$ resampled values, Mielke et al. found $T_\text{o} = 0.8420$ with an upper-tail resampling-approximation probability value of 0.0359.

6.23 Knijnenburg et al. and Probability Values

In 2009 Knijnenburg, Wessels, Reinders, and Shmulevich addressed the same problem as Brusco, Stahl, and Steinley [226]; viz., resampling permutation methods ultimately depend on the minimal obtainable probability value and the resolution of the probability value to the number of permutations [761]. Put more succinctly, for n random samples the resolution of obtainable resampling probability values is $1/n$ and the smallest achievable probability value is $1/n$. This means that a very large number of permutations is required to accurately estimate a very small probability value. To this end, Knijnenburg et al. developed a method of computing probability values based on a tail approximation using a generalized Pareto distribution (GPD). As described by Knijnenburg et al., the GPD has a cumulative distribution function (cdf) given by

$$F(z) = \begin{cases} 1 - (1 - kz/a)^{1/k} & \text{if } k \neq 0, \\ 1 - \exp(-z/a) & \text{if } k = 0, \end{cases}$$

and probability density function given by

$$f(z) = \begin{cases} a^{-1}(1 - kz/a)^{1/k-1} & \text{if } k \neq 0, \\ a^{-1}\exp(-za) & \text{if } k = 0, \end{cases}$$

where a and k are the scale and shape parameters of the Pareto distribution, respectively. The range of z is $0 \leq z < \infty$ for $k \leq 0$ and $0 \leq z \leq a/k$ for $k > 0$. For shape parameters $k = 0$ and $k = 1$, the GPD becomes the exponential and uniform distributions, respectively [761, p. i162].

Knijnenburg et al. examined seven distribution functions, ranging from light-tailed to heavy-tailed: Poisson, normal, chi-squared, exponential, F, log-normal, and Cauchy. They found in all cases, tail estimation using the GPD required fewer permutations than the standard empirical cumulative distribution function (ECDF). For probability values that were not too small (e.g., 10^{-3} to 10^{-5}) about 5 to 10 times fewer permutations were required with the GPD than with the ECDF. Finally, for the GPD approximation they observed that heavy-tailed distributions converged with fewer permutations than light-tailed distributions [761].

6.24 Reiss et al. and Multivariate Analysis of Variance

Multivariate permutation methods have found wide acceptance in ecological studies as the type of data collected usually does not satisfy the multivariate normality assumption of tests such as Pillai's Trace, Wilks' likelihood-ratio test (Λ), or Roy's Maximum Root. In 2010 Reiss, Stevens, Shehzad, Petkova, and Milham compared MRPP and pseudo-F tests in a multivariate analysis of variance context [1163]. The two methods both begin as an $n \times n$ symmetric matrix $\mathbf{D} = (d_{ij})$ for $1 \leq i, j \leq n$ representing non-negative distances among n observations. For MRPP, consider n observations divided among g a priori groups $\mathcal{G}_1, \ldots, \mathcal{G}_g$ of sizes $n_1, \ldots, n_g$, then the MRPP statistic is given by

$$\delta = \sum_{k=1}^{g} C_k \frac{2}{n_k(n_k-1)} \sum_{i<j,\, i,j \in \mathcal{G}} \Delta_{i,j} ,$$

where $\Delta_{i,j}$ denotes a measure of dissimilarity between the ith and jth observations, and

$$C_k = \frac{n_k - 1}{n - g} .$$

Following the notation of Reiss et al., for the pseudo-F test, let $\mathbf{A} = (-\frac{1}{2} d_{ij}^2)$ for $1 \leq i, j \leq n$, and let

$$\mathbf{G} = (\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)\mathbf{A}(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n),$$

where $\mathbf{1}$ is a vector of n 1s. Consider three partial design matrices, $\mathbf{X}_k$ for $k = 0, 1, 2$, where $\mathbf{X}_k$ is a $n \times m_k$ matrix of rank m_k. For $k = 0, 1, 2$, let

$$\mathbf{H}_k = \mathbf{X}_k(\mathbf{X}_k^{\mathrm{T}} \mathbf{X}_k)^{-1} \mathbf{X}_k^{\mathrm{T}}$$

be the hat matrix associated with $\mathbf{X}_k$. The pseudo-F statistic is then given by

$$F^* = \frac{\operatorname{trace}\left(\mathbf{H}_2 \mathbf{G} \mathbf{H}_2\right)/m_2}{\operatorname{trace}\left[(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})\right]/(n-m)} .$$

Thus, the pseudo-F statistic is a generalization of the classical Snedecor F test statistic that can be calculated directly from the distance matrix, whether or not the distance measurements are Euclidean. As Reiss et al. noted, this property is especially helpful in the field of ecology, where analyses often employ non-Euclidean distance measures such as that defined by Bray and Curtis in 1957 [208]:

$$d(\mathbf{x},\mathbf{y}) = \frac{\sum_{k=1}^{p} |x_k - y_k|}{\sum_{k=1}^{p} (x_k + y_k)} ,$$

where **x** and **y** are p-dimensional vectors of non-negative numbers.

Reiss et al. showed that a pseudo-F test with distance d_{ij} is equivalent to an MRPP test with dissimilarity $\Delta_{i,j} = d_{ij}^2$ and weights $C_k = (n_k - 1)/(n - g)$, where the relationship is given by

$$\delta = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2}{n\left[n - g + (g-1)F^*\right]} .$$

Much of the remainder of the paper is a comparison and evaluation of distances d_{ij}^2 and d_{ij}, the second favored by Mielke for its congruence with the data space [941]. Finally, Reiss et al. noted that there appears to be little recognition that MRPP and pseudo-F are related, which contributes to a lack of understanding across disciplines. The authors expressed hope that the equivalence presented in the paper would help to reduce "this mutual incomprehension" [1163, p. 642].

6.25 A Permutation Analysis of Trend

In 2011 Berry, Johnston, and Mielke developed a permutation alternative to the F test for the analysis of trend [116]. As Berry et al. explained, it is sometimes necessary to compare the means of treatment groups when the independent variable is quantitative. In such cases, it is more informative to consider the overall trend among the treatment groups than simply to make specific comparisons among the treatment means [116, p. 247].

The requisite F-ratios for an analysis of trend are obtained by polynomial multiple regression, where a single independent variable is raised to successive powers, i.e., $1, \ldots, k-1$. Let n_j denote the number of subjects in the jth of k treatments, $j = 1, \ldots, k$,

$$N = \sum_{j=1}^{k} n_j ,$$

let y_i, $i = 1, \ldots, N$, denote univariate measurements on the N subjects, and let x_i denote the quantitative values associated with the jth of k treatments for the ith subject, $i = 1, \ldots, N$. Then,

$$R^2_{\text{Linear}} = R^2_{y.x}\,, \quad R^2_{\text{Quadratic}} = R^2_{y.x,x^2}\,, \quad R^2_{\text{Cubic}} = R^2_{y.x,x^2,x^3}\,,$$

and so on. The F-ratio statistics are then given by

$$F_{\text{Linear}} = \frac{R^2_{y.x}}{\left(1 - R^2_{y.x,x^2,\ldots,x^{k-1}}\right) \Big/ \left(N - k\right)}\,, \tag{6.5}$$

$$F_{\text{Quadratic}} = \frac{R^2_{y.x,x^2}}{\left(1 - R^2_{y.x,x^2,\ldots,x^{k-1}}\right) \Big/ \left(N - k\right)}\,, \tag{6.6}$$

$$F_{\text{Cubic}} = \frac{R^2_{y.x,x^2,x^3}}{\left(1 - R^2_{y.x,x^2,\ldots,x^{k-1}}\right) \Big/ \left(N - k\right)}\,, \tag{6.7}$$

and so on.

F_{Linear}, $F_{\text{Quadratic}}$, and F_{Cubic} in Eqs. (6.5)–(6.7) are simply tests of significance for the appropriate squared semi-partial correlation coefficients in which the numerators of Eqs. (6.5)–(6.7) are squared semi-partial correlation coefficients and the common denominator in Eqs. (6.5)–(6.7) is the MS_{Residual}. Under the null hypothesis, the distribution of each F-ratio is Snedecor's F distribution with 1 and $N - k$ degrees of freedom. As Berry et al. noted, tests of significance for squared semi-partial correlation coefficients are especially sensitive to deviations from the assumptions of normality and homogeneity; see for example, two articles by Algina, Keselman, and Penfield in 2007 and 2010 [11, 12].

The permutation approach to the analysis of trend as developed by Berry et al. followed the conventional approach up to, but not including, the determination of the probability value. For data of this type, the determination of an exact permutation probability value is unrealistic, as the number of permutations given by

$$M = \frac{N!}{\displaystyle\prod_{j=1}^{k} n_j!}$$

is usually very large, precluding calculation of an exact probability value. Therefore, an approximate two-sided resampling probability value was obtained by computing an F-ratio on the observed data, randomly shuffling the N responses L times, redistributing the shuffled responses to the k treatments with n_j, $j = 1, \ldots, k$ held constant, computing L resampled values of F, and finding the proportion of the L resampled F-ratio values equal to or greater than the observed F-ratio value. In example analyses, Berry et al. set $L = 1{,}000{,}000$.

In comparison analyses of the conventional F test for the analysis of trend and the permutation alternative analysis of trend, Berry et al. found substantial differences when extreme values were included in the data sets. They concluded that, in these cases, the assumption of homogeneity of variance was not met and the use of the conventional F test for the analysis of trend provided erroneous results. They established that the permutation alternative to the conventional F test for the analysis of trend provided probability values that were free of the restrictive assumptions of normality and homogeneity underlying the use of the Snedecor F distribution [116, p. 254].

6.26 Curran-Everett and Permutation Methods

In 2012 Douglas Curran-Everett published an overview article in *Advances in Physiology Education* in an attempt to introduce permutation methods to researchers and students in the field of physiology [307].[39] Written in a conversational style, Curran-Everett provided a brief history of permutation methods, an overview comparing the Neyman–Pearson population and Fisher permutation models (q.v. page 3), an example based on two independent samples, a second example based on simple bivariate correlation, and a practical approach to permutation methods that is worth summarizing for its succinctness and clarity [307, p. 185].

1. Define the problem—the null hypothesis—we care about.
2. Calculate a sample statistic that is relevant to the null hypothesis.
3. Rearrange the observations in ways that are consistent with the null hypothesis.
4. For each arrangement, calculate the sample statistic.
5. Compute the proportion of sample statistics in the permutation distribution that are as or more extreme than the value of the observed sample statistic value.

Curran-Everett pointed out that when such notable researchers such as John Tukey [1382], Bradley Efron and Rob Tibshirani [402], Michael Ernst [413], Phillip Good [531], Oscar Kempthorne [719], and John Ludbrook and Hugh Dudley [856] endorse permutation methods, it is incumbent on other researchers to pay attention [307, p. 186]. In addition, he advised using permutation methods for the actual statistical analysis whenever possible to assess whether a statistical inference made from a more traditional hypothesis test is justified. Finally, Curran-Everett concluded that if the conclusion from permutation methods matches the conclusion from the traditional test of hypothesis, then one can be assured that the assumptions for the traditional procedure have been reasonably well met, citing works by Eugene Edgington and Patrick Onghena in 2007 [396], R.A. Fisher in 1960 [461], Phillip Good in 2005 [531], and Bryan Manly in 2007 [877].

[39] This was the eighth article by Curran-Everett in a series published under the rubric "Explorations in statistics" in *Advances in Physiology Education*; the previous articles in the series covered standard deviations and standard errors, confidence intervals, hypothesis tests, the bootstrap, correlation, power, and regression, and were published in 2008, 2009, 2010, and 2011 [300–306].

Epilogue

Originally developed to test and confirm the robustness of classical statistical tests and measures such as Student's t test, bivariate correlation and regression, analyses of variance for completely randomized and randomized block designs, and chi-squared tests of independence and goodness-of-fit, permutation methods have emerged as an area of statistical analysis in their own right. Presently permutation tests constitute a gold standard against which conventional tests are often evaluated; see for example, discussions by Scheffé in 1959 [1232, p. 82]; Kempthorne in 1966 and 1977 [720, 721]; Bradley in 1968 [201, p. 85]; Read and Cressie in 1988 [1157]; Bakeman, Robinson, and Quera in 1996 [50]; and Edgington and Onghena in 2007 [396, p. 9]. From their inception, permutation tests were understood by many researchers to be superior to conventional tests as permutation tests were data-dependent, did not depend on the assumptions associated with classical tests, were appropriate for use with either an entire population or a nonrandom sample, and provided exact probability values.

The fact that permutation tests yield exact probability values is still extremely important in validating conventional tests. For example, in 2000 Bergmann, Ludbrook, and Spooren [100] investigated the efficacy of a variety of statistical packages and calculated probability values for the Wilcoxon–Mann–Whitney two-sample rank-sum test. Utilizing a single data set, the probability value of the Wilcoxon–Mann–Whitney test was calculated using eleven standard statistical packages, producing a variety of different probability values. Bergmann et al. concluded that the only accurate form of the Wilcoxon–Mann–Whitney test was "one in which the exact permutation null distribution [was] compiled for the actual data" [100, p. 72] (q.v. page 171). The editor of *The American Statistician* at that time, Joseph Hilbe, further noted that "it is a cause of considerable concern when the results for a relatively simple test differ across [statistical] packages" [617, p. 71].

The Fisher exact probability test is the iconic permutation test and is familiar to most researchers. The test was introduced by Fisher in an invited paper on "The logic of inductive inference" at the annual Christmas meeting of the Royal Statistical Society on 18 December 1934, a paper that appeared in *Journal of the Royal Statistical Society* the following year [452], although the origin of the Fisher exact probability test is usually attributed to the celebrated "lady tasting tea" experiment

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9,

© Springer International Publishing Switzerland 2014

Fig. 1 Notation for a 2 × 2 contingency table

x	$r-x$	r
$c-x$	$n-r-c+x$	$n-r$
c	$n-c$	n

at the Rothamsted Experimental Station in the 1920s (q.v. page 58). In this seminal paper, Fisher analyzed a set of data from Johannes Lange on convictions of same-sex twins of criminals (q.v. page 41). A search of the web in May 2013 for "Fisher exact test" yielded 4,740,000 results. Thus, the Fisher exact probability test provides a recognizable vehicle to summarize the attributes that distinguish permutation tests from conventional tests in general. However, the test is not always used without misrepresentation or controversy; see [855, 915, 1251].

To describe the Fisher exact probability test, consider a 2 × 2 contingency table of n cases, where x denotes the frequency of any cell and r and c represent the row and column marginal frequency totals, respectively, corresponding to x; such as in Fig. 1. Given fixed marginal frequency totals, the point-probability value of x is equivalent to the point-probability value of the observed table and Fisher's exact probability value is the hypergeometric point-probability of x given by

$$p(x\,|\,n,r,c) = \frac{\dbinom{r}{x}\dbinom{n-r}{c-x}}{\dbinom{n}{c}} = \frac{r!\,c!\,(n-r)!\,(n-c)!}{n!\,x!\,(r-x)!\,(c-x)!\,(n-r-c+x)!}\ . \tag{1}$$

This, of course, is exactly the formulation of the lady tasting tea experiment; see discussions by Fisher [451, pp. 11–29], Box [195, pp. 134–135], Okamoto [1053], Salsburg [1218, pp. 1–2], Senn [1250–1252], and Springate [1313].

The probability of the observed table or one more extreme requires the enumerated permutation distribution of $a \le x \le b$, where $a = \max(0, r + c - n)$ and $b = \min(r, c)$, in the notation of Fig. 1. If x_{o} denotes the observed value of x, the point-probability value of x_{o} must first be determined, as in Eq. (1). The exact hypergeometric cumulative-probability value is then given by

$$P(x_{\text{o}}|n,r,c) = \sum_{k=a}^{b} G_k\ p(k|n,r,c)\ , \tag{2}$$

where

$$G_k = \begin{cases} 1 & \text{if } p(k|n,r,c) \le p(x_{\text{o}}|n,r,c)\ , \\ 0 & \text{otherwise}\ , \end{cases}$$

for $a \le x_{\text{o}} \le b$.

Table 1 Five possible arrangements of cell frequencies with $n = 8$ and identical marginal frequency totals of 4, 4, 4, and 4

Table 1		Table 2		Table 3		Table 4		Table 5	
0	4	1	3	2	2	3	1	4	0
4	0	3	1	2	2	1	3	0	4

The Fisher exact probability test can be illustrated by analyzing the data from the lady tasting tea experiment. Recall that the experiment consisted of preparing eight cups of tea with milk, four with the milk poured into the cup first and the tea infusion added second, four with the tea infusion poured into the cup first and the milk added second, and presenting them to the subject for judgment in random order (q.v. page 58). The subject was told in advance of what the test would consist; namely, she would be asked to taste eight cups of tea, that these would be four of each kind, and that the cups of tea would be presented to her in random order. Her task was to divide the eight cups of tea into two sets of four each, agreeing, if possible, with the treatments administered [451, Sect. 11].

For these data, $r = c = 4, n = 8, a = \max(0, r+c-n) = \max(0, 4+4-8) = 0$ and $b = \min(r, c) = \min(4, 4) = 4$. Thus, there are five possible arrangements of the data wherein $0 \leq x \leq 4$. Table 1 lists the five possible 2×2 tables with $n = 8$ and identical marginal frequency totals of $\{4, 4\}$ and $\{4, 4\}$. As in Eqs. (1) and (2), let lower-case p denote the hypergeometric point-probability value and upper-case P denote the hypergeometric cumulative-probability value of a given x. Then, following Eq. (1), the point-probability values for $x = 0, \ldots, 4$ are given by

$$p(x = 0|8, 4, 4) = \frac{4!\ 4!\ (8-4)!\ (8-4!)}{8!\ 0!\ (4-0)!\ (4-0)!\ (8-4-4+0)!} = \frac{1}{70} = 0.0143\ ,$$

$$p(x = 1|8, 4, 4) = \frac{4!\ 4!\ (8-4)!\ (8-4!)}{8!\ 1!\ (4-1)!\ (4-1)!\ (8-4-4+1)!} = \frac{16}{70} = 0.2286\ ,$$

$$p(x = 2|8, 4, 4) = \frac{4!\ 4!\ (8-4)!\ (8-4!)}{8!\ 2!\ (4-2)!\ (4-2)!\ (8-4-4+2)!} = \frac{36}{70} = 0.5153\ ,$$

$$p(x = 3|8, 4, 4) = \frac{4!\ 4!\ (8-4)!\ (8-4!)}{8!\ 3!\ (4-3)!\ (4-3)!\ (8-4-4+3)!} = \frac{16}{70} = 0.2286\ ,$$

and

$$p(x = 4|8, 4, 4) = \frac{4!\ 4!\ (8-4)!\ (8-4!)}{8!\ 4!\ (4-4)!\ (4-4)!\ (8-4-4+4)!} = \frac{1}{70} = 0.0143\ .$$

Suppose, for example, that $x_o = 3$. Then the hypergeometric cumulative probability value is the sum of the probability values less than or equal to $p(3|8, 4, 4) = 0.2286$, i.e.,

$$P = 0.0143 + 0.2286 + 0.2286 + 0.0143 = 0.4858\ .$$

Here, the exact cumulative probability value is calculated using Irwin's rule, not the doubling rule, which in this case would yield the same cumulative probability value since the marginal distributions of $\{4, 4\}$ and $\{4, 4\}$ are identical and, consequently, the discrete permutation distribution is symmetric. The controversy over the two rules erupted primarily in the 1980s (q.v. page 51); however, the argument over which rule is proper persists today. Recent articles by Dupont in 1986 and 1989 [364, 365], Lloyd in 1988 [838], Martín Andrés and Luna del Castillo in 1989 [900], and Neuhäuser in 2004 [1031] continue the controversy.

Obviously, the Fisher exact probability test is not included in the traditional Neyman–Pearson [1035, 1036] population model of conditional assignment [663, 664, 855]. Indeed, the Fisher and Neyman–Pearson approaches represent two different visions of science (q.v. page 3); see for example, a discussion by Goodman in 1993 [539]. There is no testable null hypothesis for the Fisher permutation model in the Neyman–Pearson sense of a posited population parameter, and no alternative hypothesis. Also, the Fisher permutation model contains no probability of type I error; no probability of type II error, and therefore no complement of the probability of type II error, i.e., power; no point estimate of a population parameter; and, consequently, no confidence limits. For Fisher, a computed probability value was a measure of evidence in a single experiment; whereas for Neyman–Pearson, a probability value was to be interpreted as a hypothetical frequency of error if the experiment was to be repeated many times. Finally, the Fisher exact probability test is completely data-dependent, makes no assumptions about a theoretical distribution, and does not require a random sample drawn from a specified population.

Early in their history, permutation methods were impractical and usually limited to the verification of conventional statistical tests. It was the advent of high-speed computing that allowed permutation methods to become practical. Permutation methods have since supplanted conventional statistical methods for a variety of research designs, and the field continues to expand as researchers design new applications for permutation methods. Presently, it appears that computing speed is sufficient for most applications of permutation methods. When combined with resampling methods and innovative algorithms, permutation tests are preferred alternatives to many conventional statistical tests.

References

1. Abdi, H., Williams, L.J.: Jackknife. In: Salkind, N. (ed.) Encyclopedia of Research Design, pp. 654–661. Sage, Thousand Oaks (2010)
2. Abramowitz, M., Stegun, I.A. (eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. No. 55 in Applied Mathematics Series. National Bureau of Standards, Washington, DC (1964)
3. Agresti, A.: Exact inference for categorical data: Recent advances and continuing controversies. Stat. Med. **20**, 2709–2722 (2001)
4. Agresti, A.: Categorical Data Analysis, 3rd edn. Wiley, New York (2002)
5. Agresti, A., Ghosh, A.: Raking kappa: Describing potential impact of marginal distributions of measures of agreement. Biometrical J. **37**, 811–820 (1995)
6. Agresti, A., Mehta, C.R., Patel, N.R.: Exact inference for contingency tables with ordered categories. J. Am. Stat. Assoc. **85**, 453–458 (1990)
7. Agresti, A., Wackerly, D.: Some exact conditional tests of independence for $R \times C$ cross-classification tables. Psychometrika **42**, 111–125 (1977)
8. Agresti, A., Wackerly, D., Boyett, J.M.: Exact conditional tests for cross-classifications: Approximation of attained significance levels. Psychometrika **44**, 75–83 (1979)
9. Ahlswede, R.: Jacob Wolfowitz (1910–1981). http://www.ifp.illinois.edu/~junchen/jacob_wolfowitz.htm (1981). Accessed 19 Jan 2012
10. Albers, D.J., Reid, C., Dantzig, G.B.: An interview with George B. Dantzig: The father of linear programming. College Math. J. **17**, 292–314 (1986)
11. Algina, J., Keselman, H.J., Penfield, R.D.: Confidence intervals for an effect size measure in multiple linear regression. Educ. Psychol. Meas. **67**, 207–218 (2007)
12. Algina, J., Keselman, H.J., Penfield, R.D.: Confidence intervals for squared semipartial correlation coefficients: The effect of nonnormality. Educ. Psychol. Meas. **70**, 926–940 (2010)
13. Allen, P.G.: Idea Man: A Memoir by the Cofounder of Microsoft. Portfolio/Penguin, New York (2011)
14. Alroy, J.: Four permutation tests for the presence of phylogenetic structure. Syst. Biol. **43**, 430–437 (1994)
15. Altman, D.G., Bland, J.M.: Measurement in medicine: The analysis of method comparison studies. Statistician **32**, 307–317 (1983)
16. Álvarez-Vaquero, F., Sanz-González, J.L.: Complexity analysis of permutation versus rank test for nonparametric radar detection. Radar Proc. Tech. Appl. II **3161**, 171–176 (1997) [Proceedings of the Society of Photo-optical Instrumentation Engineers]
17. Anderson, E.: The Irises of the Gaspé peninsular. Bull. Am. Iris Soc. **59**, 2–5 (1935)
18. Anderson, E.: The species problem in Iris. Ann. Mo. Bot. Gdn. **23**, 457–509 (1936)
19. Anderson, M.J., ter Braak, C.J.F.: Permutation tests for multi-factorial analysis of variance. J. Stat. Comput. Simul. **73**, 85–113 (2003)
20. Anderson, M.J., Legendre, P.: An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. J. Stat. Comput. Simul. **62**, 271–303 (1999)

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*, 433

DOI 10.1007/978-3-319-02744-9,
© Springer International Publishing Switzerland 2014

21. Anderson, M.J., Robinson, J.: Permutation tests for linear models. Aust. N. Z. J. Stat. **43**, 75–88 (2001)
22. Andriani, P., McKelvey, B.: Perspective — from Gaussian to Paretian thinking: Causes and implications of power laws in organizations. Organ. Sci. **20**, 1053–1071 (2009)
23. Anonymous: Dr P. H. Leslie. Nature **239**, 477–478 (1972)
24. Anonymous: Former RSS honorary secretary, Sidney Rosenbaum, dies. RSSeNews. http://www.rssenews.org.uk/2013/03/former-rss-honorary-secretary-sidney-rosenbaum-dies (20 March 2013). Accessed 9 June 2013
25. Anonymous: Turing top secret. Significance **9**, 3 (June 2012)
26. Ansari, A.R., Bradley, R.A.: Rank sum tests for dispersion. Ann. Math. Stat. **31**, 1174–1189 (1960)
27. Anscombe, F.J.: Rejection of outliers. Technometrics **2**, 123–147 (1960)
28. Appleby, J., Hunt, L., Jacob, M.: Telling the Truth About History. Norton, New York (1994)
29. Arboretti Giancristofaro, R., Bonnini, S., Pesarin, F.: A permutation approach for testing heterogeneity in two-sample categorical variables. Stat. Comput. **19**, 209–216 (2009)
30. Arbuckle, J., Aiken, L.S.: A program for Pitman's permutation test for differences in location. Behav. Res. Methods Instrum. **7**, 381 (1975)
31. Armitage, P.: Joseph Oscar Irwin, 1898–1982. J. R. Stat. Soc. A Gen., 526–528 (1982)
32. Armitage, P.: Joseph Oscar Irwin. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 472–474. Springer, New York (2001)
33. Armitage, P., Berry, G.: Stat. Methods Med. Res., 3rd edn. Blackwell Scientific, Oxford (1987)
34. Armsen, P.: Tables for significance tests of 2×2 contingency tables. Biometrika **42**, 494–511 (1955)
35. Arnold, H.J.: Permutation support for multivariate techniques. Biometrika **51**, 65–70 (1964)
36. Aroian, L.J.: A study of R. A. Fisher's z distribution and the related F distribution. Ann. Math. Stat. **12**, 429–448 (1941)
37. Arrow, K.J., Lehmann, E.L.: Harold Hotelling 1895–1973. Natl. Acad. Sci. Bio. Mem. **87**, 1–15 (2005)
38. Asimov, N.: Erich L. Lehmann — Berkeley professor — dies. San Francisco Chronicle, pp. D–8. http://www.sfgate.com/science/article/Erich-L-Lehmann-Berkeley-professor-dies-3214326.php (16 October 2009). Accessed 9 Sept 2010
39. Askey, R.: The 1839 paper on permutations: Its relation to the Rodrigues formula and further developments. In: Altman, S., Ortiz E.L. (eds.) Mathematics and Social Utopias in France: Olinde Rodrigues and His Times, vol. 28, History of Mathematics, pp. 105–118. American Mathematical Society, Providence (2005)
40. Auble, D.: Extended tables for the Mann–Whitney statistic. B. Inst. Educ. Res. Ind. **1**, 1–39 (1953)
41. Augarten, S.: Bit by Bit: An Illustrated History of Computers. Ticknor & Fields, New York. http://ds.haverford.edu/bitbybit/bit-by-bit-contents/chapter-five/5-8-the-ias-computer/ (1984). Accessed 7 Apr 2013
42. Babington Smith, C.: Evidence in Camera: The Story of Photographic Intelligence in World War II. David & Charles, London (1957)
43. Bacaër, N.: The Leslie matrix. In: A Short History of Mathematical Population Dynamics, chap. 21, pp. 117–120. Springer, London (2011)
44. Backus, J.: The history of FORTRAN I, II, and III. ACM SIGPLAN Notices **13**, 165–180 (1978)
45. Baglivo, J., Olivier, D., Pagano, M.: Methods for the analysis of contingency tables with large and small cell counts. J. Am. Stat. Assoc. **83**, 106–113 (1988)
46. Baglivo, J., Olivier, D., Pagano, M.: Methods for exact goodness-of-fit tests. J. Am. Stat. Assoc. **82**, 464–469 (1992)
47. Baglivo, J., Olivier, D., Pagano, M.: Analysis of discrete data: Rerandomization methods and complexity. Comp. Stat. Data. Anal. **16**, 175–184 (1993)

48. Bailer, A.J.: Testing variance equality with randomization tests. J. Stat. Comput. Simul. **31**, 1–8 (1989)
49. Bailey, R.A.: Restricted randomization: A practical example. J. Am. Stat. Assoc. **82**, 712–719 (1987)
50. Bakeman, R., Robinson, B.F., Quera, V.: Testing sequential association: Estimating exact p values using sampled permutations. Psychol. Methods **1**, 4–15 (1996)
51. Baker, F.B., Collier, Jr., R.O.: Monte Carlo F-II: A computer program for analysis of variance F-tests by means of permutation. Educ. Psychol. Meas. **26**, 169–173 (1966)
52. Baker, F.B., Collier, Jr., R.O.: Some empirical results on variance ratios under permutation in the completely randomized design. J. Am. Stat. Assoc. **61**, 813–820 (1966)
53. Baker, F.B., Hubert, L.J.: Inference procedures for ordering theory. J. Educ. Stat. **2**, 217–232 (1977)
54. Baker, R.D., Tilbury, J.B.: Algorithm 283: Rapid computation of the permutation paired and grouped t-tests. J. R. Stat. Soc. C App. **42**, 432–441 (1993)
55. Baker, R.J.: Algorithm 112: Exact distributions derived from two-way tables. J. R. Stat. Soc. C Appl. Stat. **26**, 199–206 (1977) [Correction: J. R. Stat. Soc. C Appl. Stat. **27**, 109 (1978)]
56. Balmer, D.W.: Algorithm 236: Recursive enumeration of $r \times c$ tables for exact likelihood evaluation. J. R. Stat. Soc. C Appl. Stat. **37**, 290–301 (1988)
57. Bancroft, T.A.: George W. Snedecor: A chronology. In: Bancroft T.A. (ed.) Statistical Papers in Honor of George W. Snedecor, pp. ix–xi. Iowa State University Press, Ames (1972)
58. Bancroft, T.A.: Highlights of some expansion years of the Iowa State Statistical Laboratory, 1947–72. In: David, H.A., David, H.T. (eds.) Statistics: An Appraisal, pp. 19–30. Iowa State University Press, Ames (1984)
59. Bancroft, T.A.: Snedecor, George Waddel. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 339–341. Wiley, New York (1997)
60. Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D.: Beyond kappa: A review of interrater agreement measures. Can. J. Stat. **27**, 3–23 (1999)
61. Barbella, P., Denby, L., Landwehr, J.M.: Beyond exploratory data analysis: The randomization test. Math. Teach. **83**, 144–149 (February 1990)
62. Barboza, D., Markoff, J.: Power in numbers: China aims for high-tech primacy. NY Times **161**, D2–D3 (6 December 2011)
63. Barnard, G.A.: A new test for 2 × 2 tables. Nature **156**, 177 (1945)
64. Barnard, G.A.: A new test for 2 × 2 tables. Nature **156**, 783–784 (1945)
65. Barnard, G.A.: 2×2 tables. A note on E. S. Pearson's paper. Biometrika **34**, 168–169 (1947)
66. Barnard, G.A.: The meaning of a significance level. Biometrika **34**, 179–182 (1947)
67. Barnard, G.A.: Significance tests for 2 × 2 tables. Biometrika **34**, 123–138 (1947)
68. Barnard, G.A.: Statistical inference. J. R. Stat. Soc. B Met. **11**, 115–149 (1949)
69. Barnard, G.A.: In contradiction to J. Berkson's dispraise: Conditional tests can be more efficient. J. Stat. Plan. Infer. **3**, 181–187 (1979)
70. Barnard, G.A.: Discussion of "Tests of significance in 2 × 2 tables" by F. Yates. J. R. Stat. Soc. A Gen. **147**, 449–450 (1984)
71. Barnard, G.A.: The early history of the Fisher–Yates–Irwin formula and Fisher's 'exact test'. J. Stat. Comput. Simul. **20**, 153–155 (1984)
72. Barnard, G.A.: Discussion of "A new probability model for determining exact P-values for 2 × 2 contingency tables when comparing binomial proportions" by W.R. Rice. Biometrics **44**, 16–18 (1988)
73. Barnard, G.A.: Kendall, Maurice George. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 130–132. Wiley, New York (1997)
74. Barrodale, I., Roberts, F.D.K.: A improved algorithm for discrete ℓ_1 linear approximation. J. Num. Anal. **10**, 839–848 (1973)
75. Barrodale, I., Roberts, F.D.K.: Solution of an overdetermined system of equations in the ℓ_1 norm. Commun. ACM **17**, 319–320 (1974)

76. Bartholomew, D.J.: Egon Sharpe Pearson. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 373–376. Springer, New York (2001)
77. Bartko, J.J.: Measurement and reliability — statistical thinking considerations. Schizophrenia Bull. **17**, 483–489 (1991)
78. Bartlett, M.S.: Properties of sufficiency and statistical tests. P. R. Soc. Lond. A Math. **160**, 268–282 (1937)
79. Bartlett, M.S.: A note on tests of significance in multivariate analysis. Proc. Camb. Philos. Soc. **34**, 33–40 (1939)
80. Bartlett, M.S.: R. A. Fisher and the last fifty years of statistical methodology. J. Am. Stat. Assoc. **60**, 395–409 (1965)
81. Bartlett, M.S.: J.O. Irwin, 1898–1982. Int. Stat. Rev. **52**, 109–114 (1984)
82. Barton, D.E., David, F.N.: A test for birth order effect. Ann. Hum. Genet. **22**, 250–257 (1957–1958)
83. Barton, D.E., David, F.N.: Randomization bases for multivariate tests I. The bivariate case: Randomness of n points in a plane. B. Int. Stat. Inst. **39**, 455–467 (1961)
84. Bassett, Jr., G., Koenker, R.: Asymptotic theory of least absolute error regression. J. Am. Stat. Assoc. **73**, 618–622 (1978)
85. Basu, D.: Discussion of Joseph Berkson's paper "In dispraise of the exact test". J. Stat. Plan. Infer. **3**, 189–192 (1979)
86. Basu, D.: Randomization analysis of experimental data: The Fisher randomization test (with discussion). J. Am. Stat. Assoc. **75**, 575–582 (1980)
87. Basu, D.: Rejoinder to comments on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. J. Am. Stat. Assoc. **75**, 593–595 (1980)
88. Bear, G.: Computationally intensive methods warrant reconsideration of pedagogy in statistics. Behav. Res. Methods Instrum. C **27**, 144–147 (1995)
89. Beaton, A.E.: Salvaging experiments: Interpreting least squares in non-random samples. In: Hogben, D., Fife, D. (eds.) Computer Science and Statistics: Tenth Annual Symposium on the Interface, pp. 137–145. U.S. Department of Commerce, Washington, DC (1978)
90. Bebbington, A.C.: A simple method of drawing a sample without replacement. J. R. Stat. Soc. C Appl. Stat. **24**, 136 (1975)
91. Bedeian, A.G., Armenakis, A.A.: A program for computing Fisher's exact probability test and the coefficient of association λ for $n \times m$ contingency tables. Educ. Psychol. Meas. **37**, 253–256 (1977)
92. Behrenz, P.G.: Algorithm 133: Random. Commun. ACM **5**, 553 (1962)
93. Bell, C.B., Sen, P.K.: Randomization procedures. In: Krishnaiah, P.R., Sen, P.K. (eds.) Nonparametric Methods, vol. IV, Handbook of Statistics, pp. 1–29. North-Holland, Amsterdam (1984)
94. Bennett, B.M., Nakamura, E.: Tables for testing significance in a 2×3 contingency table. Technometrics **5**, 501–511 (1963)
95. Bennett, E.M., Alpert, R., Goldstein, A.C.: Communications through limited-response questioning. Public Opin. Quart. **18**, 303–308 (1954)
96. Bennett, J.H. (ed.): Natural Selection, Heredity and Eugenics: Selected Correspondence of R.A. Fisher with Leonard Darwin and Others. Clarendon, Oxford (1983)
97. Bennett, J.H. (ed.): Statistical Inference and Analysis: Selected Correspondence of R.A. Fisher. Clarendon, Oxford (1990)
98. Berger, V.W.: Comment on "Why permutation tests are superior to t and F tests in biomedical research" by J. Ludbrook and H.A.F. Dudley. Am. Stat. **54**, 85–86 (2000)
99. Berger, V.W.: Pros and cons of permutation tests in clinical trials. Stat. Med. **19**, 1319–1328 (2000)
100. Bergmann, R., Ludbrook, J., Spooren, W.P.J.M.: Different outcomes of the Wilcoxon–Mann–Whitney test from different statistics packages. Am. Stat. **54**, 72–77 (2000)
101. Berkson, J.: Do the marginal totals of the 2×2 table contain relevant information respecting the table proportions? J. Stat. Plan. Infer. **2**, 43–44 (1978)

102. Berkson, J.: In dispraise of the exact test: Do the marginal totals of the 2 × 2 table contain relevant information respecting the table proportions? J. Stat. Plan. Infer. **2**, 27–42 (1978)
103. Berlin, L.: The Man Behind the Microchip: Robert Noyce and the Invention of Silicon Valley. Oxford University Press, New York (2005)
104. Bernardin, H.J., Beatty, R.W.: Performance Appraisal: Assessing Human Behavior at Work. Kent, Boston (1984)
105. Bernoulli, D.: Indicatio maxime probabilis plurium observationum discrepantium atque verisimilluma inductio inde formanda (The most probable choice between several discrepant observations and the formation therefrom of the most likely induction). Acta Acad. Sci. Petropol. **1**, 1–33 (1777) [See the English translation by C.G. Allen in Biometrika **48**, 1–18 (1961)]
106. Bernstein, J.: The Analytical Engine: Computers, Past, Present, and Future. Random House, New York (1964)
107. Berry, G., Armitage, P.: Mid-*P* confidence intervals: A brief review. Statistician **44**, 417–423 (1995)
108. Berry, K.J.: Algorithm 179: Enumeration of all permutations of multi-sets with fixed repetition numbers. J. R. Stat. Soc. C Appl. Stat. **31**, 169–173 (1982)
109. Berry, K.J.: A generator for permutations with fixed repetitions. APL Quote Quad **17**, 28 (1987)
110. Berry, K.J.: An APL function for the asymptotic test of significance for Goodman and Kruskal's gamma statistic. Behav. Res. Methods Instrum. C **21**, 473–476 (1989)
111. Berry, K.J., Johnston, J.E., Mielke, P.W.: Exact goodness-of-fit tests for unordered equiprobable categories. Percept. Motor Skill **98**, 909–918 (2004)
112. Berry, K.J., Johnston, J.E., Mielke, P.W.: Exact and resampling probability values for weighted kappa. Psychol. Rep. **96**, 243–252 (2005)
113. Berry, K.J., Johnston, J.E., Mielke, P.W.: Exact and resampling probability values for measures associated with ordered *R* by *C* contingency tables. Psychol. Rep. **99**, 231–238 (2006)
114. Berry, K.J., Johnston, J.E., Mielke, P.W.: Exact permutation probability values for weighted kappa. Psychol. Rep. **102**, 53–57 (2008)
115. Berry, K.J., Johnston, J.E., Mielke, P.W.: Weighted kappa for multiple raters. Percept. Motor Skill. **107**, 837–848 (2008)
116. Berry, K.J., Johnston, J.E., Mielke, P.W.: Analysis of trend: A permutation alternative to the *F* test. Percept. Motor Skill. **112**, 247–257 (2011)
117. Berry, K.J., Johnston, J.E., Mielke, P.W.: Permutation methods. Comput. Stat. **3**, 527–542 (2011)
118. Berry, K.J., Kvamme, K.L., Mielke, P.W.: A permutation technique for the spatial analysis of artifacts into classes. Am. Antiquity **45**, 55–59 (1980)
119. Berry, K.J., Kvamme, K.L., Mielke, P.W.: Improvements in the permutation test for the spatial analysis of the distribution of artifacts into classes. Am. Antiquity **48**, 547–553 (1983)
120. Berry, K.J., Mielke, P.W.: Computation of finite population parameters and approximate probability values for multi-response permutation procedures (MRPP). Commun. Stat. Simul. C **12**, 83–107 (1983)
121. Berry, K.J., Mielke, P.W.: Moment approximations as an alternative to the *F* test in analysis of variance. Br. J. Math. Stat. Psychol. **36**, 202–206 (1983)
122. Berry, K.J., Mielke, P.W.: A rapid FORTRAN subroutine for the Fisher exact probability test. Educ. Psychol. Meas. **43**, 167–171 (1983)
123. Berry, K.J., Mielke, P.W.: Computation of exact probability values for multi-response permutation procedures (MRPP). Commun. Stat. Simul. C **13**, 417–432 (1984)
124. Berry, K.J., Mielke, P.W.: An APL function for Radlow and Alf's exact chi-square test. Behav. Res. Methods Instrum C **17**, 131–132 (1985)
125. Berry, K.J., Mielke, P.W.: Computation of exact and approximate probability values for a matched-pairs permutation test. Commun. Stat. Simul. C. **14**, 229–248 (1985)

126. Berry, K.J., Mielke, P.W.: Goodman and Kruskal's tau-b statistic: A nonasymptotic test of significance. Sociol. Methods Res. **13**, 543–550 (1985)
127. Berry, K.J., Mielke, P.W.: Subroutines for computing exact chi-square and Fisher's exact probability tests. Educ. Psychol. Meas. **45**, 153–159 (1985)
128. Berry, K.J., Mielke, P.W.: An APL function for computing measures of association for nominal-by-ordinal and ordinal-by-ordinal cross classifications. Behav. Res. Methods Instrum. C **18**, 399–402 (1986)
129. Berry, K.J., Mielke, P.W.: *R* by *C* chi-square analyses of small expected cell frequencies. Educ. Psychol. Meas. **46**, 169–173 (1986)
130. Berry, K.J., Mielke, P.W.: Goodman and Kruskal's tau-b statistic: A FORTRAN 77 subroutine. Educ. Psychol. Meas. **46**, 646–649 (1986)
131. Berry, K.J., Mielke, P.W.: Exact chi-square and Fisher's exact probability test for 3 by 2 cross-classification tables. Educ. Psychol. Meas. **47**, 631–636 (1987)
132. Berry, K.J., Mielke, P.W.: APL approximations for common statistical critical values. Behav. Res. Methods Instrum. C **20**, 339–342 (1988)
133. Berry, K.J., Mielke, P.W.: A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ. Psychol. Meas. **48**, 921–933 (1988)
134. Berry, K.J., Mielke, P.W.: Monte Carlo comparisons of the asymptotic chi-square and likelihood-ratio tests with the nonasymptotic chi-square test for sparse *R* by *C* tables. Psychol. Bull. **103**, 256–264 (1988)
135. Berry, K.J., Mielke, P.W.: Simulated power comparisons of the asymptotic and nonasymptotic Goodman and Kruskal tau tests for sparse *R* by *C* tables. In: Srivastava, J.N. (ed.) Probability and Statistics: Essays in Honor of Franklin A. Graybill, pp. 9–19. North-Holland, Amsterdam (1988)
136. Berry, K.J., Mielke, P.W.: Analyzing independence in *r*-way contingency tables. Educ. Psychol. Meas. **49**, 605–607 (1989)
137. Berry, K.J., Mielke, P.W.: A generalized agreement measure. Educ. Psychol. Meas. **50**, 123–125 (1990)
138. Berry, K.J., Mielke, P.W.: A family of multivariate measures of association for nominal independent variables. Educ. Psychol. Meas. **52**, 41–55 (1992)
139. Berry, K.J., Mielke, P.W.: A measure of association for nominal independent variables. Educ. Psychol. Meas. **52**, 895–898 (1992)
140. Berry, K.J., Mielke, P.W.: Nonasymptotic goodness-of-fit tests for categorical data. Educ. Psychol. Meas. **54**, 676–679 (1994)
141. Berry, K.J., Mielke, P.W.: Exact cumulative probabilities for the multinomial distribution. Educ. Psychol. Meas. **55**, 769–772 (1995)
142. Berry, K.J., Mielke, P.W.: Analysis of multivariate matched-pairs data: A FORTRAN 77 program. Percept. Motor Skill. **83**, 788–790 (1996)
143. Berry, K.J., Mielke, P.W.: Nonasymptotic probability values for Cochran's *Q* statistic: A FORTRAN 77 program. Percept. Motor Skill. **82**, 303–306 (1996)
144. Berry, K.J., Mielke, P.W.: Agreement measure comparisons between two independent sets of raters. Educ. Psychol. Meas. **57**, 360–364 (1997)
145. Berry, K.J., Mielke, P.W.: Exact and approximate probability values for the Terpstra–Jonckheere test against ordered alternatives. Percept. Motor Skill. **85**, 107–111 (1997)
146. Berry, K.J., Mielke, P.W.: Measuring the joint agreement between multiple raters and a standard. Educ. Psychol. Meas. **57**, 527–530 (1997)
147. Berry, K.J., Mielke, P.W.: Spearman's footrule as a measure of agreement. Psychol. Rep. **80**, 839–846 (1997)
148. Berry, K.J., Mielke, P.W.: Extension of Spearman's footrule to multiple rankings. Psychol. Rep. **82**, 376–378 (1998)
149. Berry, K.J., Mielke, P.W.: A FORTRAN program for permutation covariate analyses of residuals based on Euclidean distance. Psychol. Rep. **82**, 371–375 (1998)
150. Berry, K.J., Mielke, P.W.: Least absolute regression residuals: Analyses of block designs. Psychol. Rep. **83**, 923–929 (1998)

151. Berry, K.J., Mielke, P.W.: Least sum of absolute deviations regression: Distance, leverage, and influence. Percept. Motor Skill. **86**, 1063–1070 (1998)
152. Berry, K.J., Mielke, P.W.: The negative hypergeometric probability distribution: Sampling without replacement from a finite population. Percept. Motor Skill. **86**, 207–210 (1998)
153. Berry, K.J., Mielke, P.W.: Least absolute regression residuals: Analyses of randomized designs. Psychol. Rep. **84**, 947–954 (1999)
154. Berry, K.J., Mielke, P.W.: Least absolute regression residuals: Analyses of split-plot designs. Psychol. Rep. **85**, 445–453 (1999)
155. Berry, K.J., Mielke, P.W.: Exact and Monte Carlo resampling procedures for the Wilcoxon–Mann–Whitney and Kruskal–Wallis tests. Percept. Motor Skill. **91**, 749–754 (2000)
156. Berry, K.J., Mielke, P.W.: A Monte Carlo investigation of the Fisher Z transformation for normal and nonnormal distributions. Psychol. Rep. **87**, 1101–1114 (2000)
157. Berry, K.J., Mielke, P.W.: Nonasymptotic significance tests for two measures of agreement. Percept. Motor Skill. **93**, 109–114 (2001)
158. Berry, K.J., Mielke, P.W., Helmericks, S.G.: Exact confidence limits for proportions. Educ. Psychol. Meas. **48**, 713–716 (1988)
159. Berry, K.J., Mielke, P.W., Helmericks, S.G.: An algorithm to generate discrete probability distributions: Binomial, hypergeometric, negative binomial, inverse hypergeometric, and Poisson. Behav. Res. Methods Instrum. C **26**, 366–367 (1994)
160. Berry, K.J., Mielke, P.W., Johnston, J.E.: The two-sample rank-sum test: early development. Elec. J. Hist. Prob. Stat. **8**, 1–26 (2012)
161. Berry, K.J., Mielke, P.W., Kvamme, K.L.: Efficient permutation procedures for analysis of artifact distributions. In: Hietala, H.J. (ed.) Intrasite Spatial Analysis in Archaeology, pp. 54–74. Cambridge University Press, Cambridge (1984)
162. Berry, K.J., Mielke, P.W., Mielke, H.W.: The Fisher–Pitman permutation test: An attractive alternative to the *F* test. Psychol. Rep. **90**, 495–502 (2002)
163. Bertrand, J.L.F.: Calcul des Probabilitiés. Gauthier-Villars et fils, Paris (1889) [Reprinted by Chelsea Publishing (AMS), New York, in 1972]
164. Besag, J., Diggle, P.J.: Simple Monte Carlo tests for spatial pattern. J. R. Stat. Soc. C Appl. Stat. **26**, 327–333 (1977)
165. Beyer, K.W.: Grace Hopper and the Invention of the Information Age. MIT Press, Cambridge (2009)
166. Biondini, M.E., Mielke, P.W., Berry, K.J.: Data-dependent permutation techniques for the analysis of ecological data. Vegetatio **75**, 161–168 (1988) [The name of the journal was changed to *Plant Ecology* in 1997]
167. Bitner, J.R., Ehrlich, G., Reingold, E.M.: Efficient generation of the binary reflected Gray code and its application. Commun. ACM **19**, 517–521 (1976)
168. Blair, R.C., Higgins, J.J.: A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* under various nonnormal distributions. J. Educ. Stat. **5**, 309–335 (1980)
169. Blair, R.C., Higgins, J.J., Karniski, W., Kromrey, J.D.: A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances. Multivar. Behav. Res. **29**, 141–163 (1994)
170. Blair, R.C., Troendle, J.F., Beck, R.W.: Control of familywise errors in multiple endpoint assessments via stepwise permutation tests. Stat. Med. **15**, 1107–1121 (1996)
171. Blattberg, R., Sargent, T.: Regression with non-Gaussian stable disturbances. Econometrica **39**, 501–510 (1971)
172. Blaug, M.: The myth of the old Poor Law and the making of the new. J. Econ. Hist. **23**, 151–184 (1963)
173. Bloomfield, P., Steiger, W.: Least absolute deviations curve-fitting. SIAM J. Sci. Stat. Comput. **1**, 290–301 (1980)
174. Boardman, T.J.: Smaller computers: Impact on statistical data analysis. In: David, H.A., David, H.T. (eds.) Statistics: An Appraisal, pp. 625–641. Iowa State University Press, Ames (1984)

175. Boik, R.J.: The Fisher–Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. Br. J. Math. Stat. Psychol. **40**, 26–42 (1987)
176. Boik, R.J.: Randomization. In: Everitt, B.S., Howell, D.C. (eds.) Encyclopedia of Statistics in Behavioral Science, vol. IV, pp. 1669–1674. Wiley, New York (2005)
177. Boland, P.J.: A biographical glimpse of William Sealy Gosset. Am. Stat. **38**, 179–183 (1984)
178. Boothroyd, J.: Algorithm 6: PERM. Comput. Bull. **3**, 104 (1965) [Reprinted in Comput. J. **22**, 88–89]
179. Boothroyd, J.: Algorithm 27: Rearrange the elements of an array section according to a permutation of the subscripts. Comput. J. **10**, 310 (1967)
180. Boothroyd, J.: Algorithm 29: Permutations of the elements of a vector. Comput. J. **10**, 310–311 (1967)
181. Boothroyd, J.: Algorithm 30: Fast permutation of the elements of a vector. Comput. J. **10**, 311–312 (1967)
182. Borgatta, E.F.: My student, the purist: A lament. Sociol. Quart. **9**, 29–34 (1968)
183. Borkowf, C.B.: An efficient algorithm for generating two-way contingency tables with fixed marginal totals and arbitrary mean proportions, with applications to permutation tests. Comput. Stat. Data Anal. **44**, 431–449 (2004)
184. Bottomley, W.B.: Sir Joseph Henry Gilbert: 1817–1901. In: Oliver, F.W. (ed.) Makers of British Botany: A Collection of Biographies by Living Botanists, pp. 233–242. Cambridge University Press, Cambridge (1913)
185. Boulton, D.M.: Remark on algorithm 434. Commun. ACM **17**, 326 (1974)
186. Boulton, D.M.: Exact probabilities for $R \times C$ contingency tables. ACM Trans. Math. Software **2**, 108 (1976)
187. Boulton, D.M., Wallace, C.S.: Occupancy of a rectangular array. Comput. J. **16**, 57–63 (1973)
188. Bowen, J.: Alan Turing. In: Robinson, A. (ed.) The Scientists: An Epic of Discovery, pp. 270–275. Thames & Hudson, London (2012)
189. Bowley, A.L.: F. Y. Edgeworth's Contributions to Mathematical Statistics. Royal Statistical Society, London (1928)
190. Box, G.E.P.: Non-normality and tests on variances. Biometrika **40**, 318–335 (1953)
191. Box, G.E.P.: Science and statistics. J. Am. Stat. Assoc. **71**, 791–799 (1976)
192. Box, G.E.P.: An Accidental Statistician: The Life and Memories of George E. P. Box. Wiley, New York (2013) [Also inscribed "With a little help from my friend, Judith L. Allen"]
193. Box, G.E.P., Andersen, S.L.: Permutation theory in the derivation of robust criteria and the study of departures from assumption (with discussion). J. R. Stat. Soc. B Met. **17**, 1–34 (1955)
194. Box, G.E.P., Hunter, W.G., Hunter, J.S.: Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. Wiley, New York (1978)
195. Box, J.F.: R. A. Fisher: The Life of a Scientist. Wiley, New York (1978)
196. Box, J.F.: Gosset, Fisher, and the t distribution. Am. Stat. **35**, 61–66 (1981)
197. Box, J.F.: Fisher, Ronald Aylmer. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 99–108. Wiley, New York (1997)
198. Boyer, G.R.: An Economic History of the English Poor Law: 1750–1850. Cambridge University Press, Cambridge (1990)
199. Boyett, J.M.: Algorithm 144: $R \times C$ tables with given row and column totals. J. R. Stat. Soc. C Appl. Stat. **28**, 329–332 (1979)
200. Bradbury, I.: Analysis of variance versus randomization — a comparison. Br. J. Math. Stat. Psychol. **40**, 177–187 (1987)
201. Bradley, J.V.: Distribution-free Statistical Tests. Prentice-Hall, Englewood Cliffs (1968)
202. Bradley, J.V.: A common situation conducive to bizarre distribution shapes. Am. Stat. **31**, 147–150 (1977)
203. Bradley, R.A.: Frank Wilcoxon. Biometrics **22**, 192–194 (1966)

204. Bradley, R.A.: Frank Wilcoxon. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 339–341. Wiley, New York (1997)
205. Bradley, R.A., Hollander, M.: Wilcoxon, Frank. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 420–424. Springer, New York (2001)
206. Bratley, P.: Algorithm 306: Permutations with repetitions. Commun. ACM **7**, 450–451 (1967)
207. Braun, H.I. (ed.): The Collected Works of John W. Tukey: Multiple Comparisons, 1949–1983, vol. VIII. Chapman & Hall, New York (1994)
208. Bray, J.R., Curtis, J.T.: An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. **27**, 326–349 (1957)
209. Brennan, P.F., Hays, B.J.: The kappa statistic for establishing interrater reliability in the secondary analysis of qualitative clinical data. Res. Nurs. Health **15**, 153–158 (1992)
210. Brennan, R.L., Prediger, D.J.: Coefficient kappa: Some uses, misuses, and alternatives. Educ. Psychol. Meas. **41**, 687–699 (1981)
211. Brillinger, D.R.: The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. Rev. Int. Stat. Inst. **32**, 202–206 (1964)
212. Brillinger, D.R. (ed.): The Collected Works of John W. Tukey: Time Series, 1949–1964, vol. I. Statistics/Probability Series. Wadsworth, Belmont (1984)
213. Brillinger, D.R. (ed.): The Collected Works of John W. Tukey: Time Series, 1965–1984, vol. II. Statistics/Probability Series. Wadsworth, Belmont (1985)
214. Brillinger, D.R.: John W. Tukey: His life and professional contributions. Ann. Stat. **30**, 1535–1575 (2002)
215. Brillinger, D.R.: Erich Leo Lehmann, 1917 – 2009. J. R. Stat. Soc. A Stat **173**, 683–689 (2010)
216. Brillinger, D.R., Jones, L.V., Tukey, J.W.: The role of statistics in weather resources management. Tech. Rep. II, Weather Modification Advisory Board, United States Department of Commerce, Washington, DC (1978)
217. Brin, S., Page, L.: Academy of Achievement. http://www.achievement.org/autodoc/page/pag0bio-1 (2011). Accessed 3 Nov 2012
218. Sergey Brin Biography. http://www.biography.com/people/sergey-brin-12103333 (2012). Accessed 3 Nov 2012
219. Brockwell, P.J., Mielke, P.W.: Asymptotic distributions of matched-pairs permutation statistics based on distance measures. Aust. J. Stat. **26**, 30–38 (1984)
220. Brockwell, P.J., Mielke, P.W., Robinson, J.: On non-normal invariance principles for multi-response permutation procedures. Aust. J. Stat. **24**, 33–41 (1982)
221. Brooks, E.B.: Frank Wilcoxon, 2 Sept 1892 – 18 Nov 1965. Tales of Statisticians. http://www.umass.edu/wsp/statistics/tales/wilcoxon.html. Accessed 1 Apr 2012
222. Bross, I.D.J.: Is there an increased risk? Fed. Proc. **13**, 815–819 (1954)
223. Bross, I.D.J.: How to use ridit analysis. Biometrics **14**, 18–38 (1958)
224. Brown, B.M., Maritz, J.S.: Distribution-free methods in regression. Aust. J. Stat. **24**, 318–331 (1982)
225. Brown, G.W.: History of RAND's random digits — Summary. In: Householder, A.S., Forsythe, G.E., Germond, H.H. (eds.) The Monte Carlo Method, no. 12 in National Bureau of Standards Applied Mathematics Series, pp. 31–32. United States Government Printing Office, Washington, DC (1951)
226. Brusco, M.J., Stahl, S., Steinley, D.: An implicit enumeration method for an exact test of weighted kappa. Br. J. Math. Stat. Psychol. **61**, 439–452 (2008)
227. Bryant, H.N.: The role of permutation tail probability tests in phylogenetic systematics. Syst. Biol. **41**, 258–263 (1992)
228. Buckles, B.P., Lybanon, M.: Algorithm 515: Generation of a vector from the lexicographical index. ACM Trans. Math. Software **3**, 180–182 (1977)

229. Burr, E.J.: The distribution of Kendall's score *S* for a pair of tied rankings. Biometrika **47**, 151–171 (1960)
230. Burt, C.: Professor Charles E. Spearman, F.R.S. Eugen. Rev. **37**, 187 (1946)
231. Byrt, T.: Problems with kappa. J. Clin. Epidemiol. **45**, 1452 (1992)
232. Byrt, T., Bishop, J., Carlin, J.B.: Bias, prevalence and kappa. J. Clin. Epidemiol. **46**, 423–429 (1993)
233. Cade, B.S., Richards, J.D.: Permutation tests for least absolute deviation regression. Biometrics **52**, 886–902 (1996)
234. Cade, B.S., Richards, J.D.: A permutation test for quantile regression. J. Agric. Biol. Environ. Sci. **11**, 106–126 (2006)
235. Cai, J.W., Shen, Y.: Permutation tests for comparing marginal survival functions with clustered failure time data. Stat. Med. **19**, 2963–2973 (2000)
236. Cai, L.: Multi-response permutation procedure as an alternative to the analysis of variance: An SPSS implementation. Behav. Res. Methods **38**, 51–59 (2006)
237. Cajori, F.: History of symbols for $\underline{n}$ = factorial. Isis **3**, 414–418 (1921)
238. Camic, C., Xie, Y.: The statistical turn in American social science: Columbia University, 1890 to 1915. Am. Sociol. Rev. **59**, 773–805 (1994)
239. Campbell, I.: Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. Stat. Med. **26**, 3661–3675 (2007)
240. Campbell-Kelly, M.: John von Neumann. In: Robinson, A. (ed.) The Scientists: An Epic of Discovery, pp. 276–279. Thames & Hudson, London (2012)
241. Cantor, A.: A computer algorithm for testing significance in $M \times K$ contingency tables. In: Fifth Proceedings of the Statistical Computing Section of the American Statistical Association, vol. 44, pp. 220–221. American Statistical Association, Washington, DC (1979)
242. Carey, G., Gottesman, I.I.: Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. Arch. Gen. Psychiatr. **35**, 1454–1459 (1978)
243. Carriquiry, A.L., David, H.A.: George Waddel Snedecor. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 346–351. Springer, New York (2001)
244. Case, L.: Intel's Ivy Bridge processor: Leaner and meaner. http://www.pcadvisor.co.uk/news/pc-components/3353194/intels-ivy-bridge-processor-leaner-meaner/ (23 April 2012). Accessed 29 Apr 2012
245. Castellan, N.J.: Shuffling arrays: Appearances may be deceiving. Behav. Res. Methods Instrum. C **24**, 72–77 (1992)
246. Cattell, R.B.: Charles Edward Spearman. In: Kruskal, W.H., Tanur, J.M. (eds.) International Encyclopedia of Statistics, vol. II, pp. 1036–1039. Free Press, New York (1978)
247. Chase, P.J.: Algorithm 382: Combinations of *M* out of *N* objects. Commun. ACM **13**, 368–369 (1970)
248. Chase, P.J.: Algorithm 383: Permutations of a set with repetitions. Commun. ACM **13**, 368–369 (1970)
249. Chase, P.J.: Remark on algorithm 382: Combinations of *M* out of *N* objects. Commun. ACM **13**, 376 (1970)
250. Chen, R.S., Dunlap, W.P.: SAS procedures for approximate randomization tests. Behav. Res. Methods Instrum. C **25**, 406–409 (1993)
251. Chen, Y.P.: Do the chi-square test and Fisher's exact test agree in determining extreme for 2 × 2 tables? Am. Stat. **65**, 239–245 (2011)
252. Chiang, C.L.: Statisticians in History: Jerzy Neyman: 1894–1981. http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=11. Accessed 16 Dec 2011
253. Chihara, L.M., Hesterberg, T.C.: Mathematical Statistics with Resampling and R. Wiley, New York (2011)
254. Chung, J.H., Fraser, D.A.S.: Randomization tests for a multivariate two-sample problem. J. Am. Stat. Assoc. **53**, 729–735 (1958)

255. Cicchetti, D., Allison, A.: A new procedure for assessing reliability of scoring EEG sleep recordings. Am. J. EEG Technol. **11**, 101–109 (1971)
256. Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. **43**, 551–558 (1990)
257. Cicchetti, D.V., Fleiss, J.L.: Comparison of the null distribution of weighted kappa and the *C* ordinal statistic. Appl. Psychol. Meas. **1**, 195–201 (1977)
258. Cleveland, W.S. (ed.): The Collected Works of John W. Tukey: Graphics, 1965–1985, vol. V. Wadsworth Statistics and Probability Series. Wadsworth & Brooks/Cole, Pacific Grove (1988)
259. Cochran, W.G.: The comparison of percentages in matched samples. Biometrika **37**, 256–266 (1950)
260. Cochran, W.G.: Some methods for strengthening the common χ^2 test. Biometrics **10**, 417–452 (1954)
261. Cochran, W.G.: Fisher and the analysis of variance. In: Feinberg, S.E., Hinkley, D.V. (eds.) R. A. Fisher: An Appreciation, pp. 17–34. Springer, Heidelberg (1980)
262. Cochrane, R., Duffy, J.: Psychology and scientific method. B. Br. Psychol. Soc. **27**, 117–121 (1974)
263. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960)
264. Cohen, J.: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. **70**, 213–220 (1968)
265. Cohen, J.: Weighted chi square: An extension of the kappa method. Educ. Psychol. Meas. **32**, 61–74 (1972)
266. Cohen, M.E., Arthur, J.S.: Randomization analysis of dental data characterized by skew and variance heterogeneity. Commun. Dent. Oral **19**, 185–189 (1991)
267. Cohen, S.: Structuralism and the writing of intellectual history. Hist. Theor. **17**, 175–206 (1978)
268. Collier, Jr., R.O., Baker, F.B.: Some Monte Carlo results on power of *F* tests under permutation in the simple randomized block design. Biometrika **53**, 199–203 (1966)
269. Collins, M.F.: A permutation test for planar regression. Aust. J. Stat. **29**, 303–308 (1987)
270. Collison, D.M.: Certification of algorithm 86 PERMUTE. Commun. ACM **5**, 440 (1962)
271. Conover, W.J.: Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables (with comments). J. Am. Stat. Assoc. **69**, 374–376 (1974)
272. Conover, W.J.: Practical Nonparametric Statistics, 3rd edn. Wiley, New York (1999)
273. Conover, W.J., Iman, R.L.: Rank transformations as a bridge between parametric and nonparametric statistics (with discussion). Am. Stat. **35**, 124–129 (1981)
274. Conroy, R.M.: What hypotheses do "nonparametric" two-group tests actually test? Stata J. **12**, 182–190 (2012)
275. Constable, S.: When investing, try thinking outside the box. http://online.wsj.com/article/SB10001424052970203960804577241263821844868.html#mod=sunday_journal_primary_hs (26 February 2012). Accessed 29 Feb 2012
276. Cooper, B.E.: Discussion of "Present position and potential developments: Some personal views" by J.A. Nelder. J. R. Stat. Soc. A Gen. **147**, 159–160 (1984)
277. Corain, L., Salmaso, L.: A critical review and a comparative study on conditional permutation tests for two-way ANOVA. Commun. Stat. Simul. C **36**, 791–805 (2007)
278. Cordell, D., Drangert, J.O., White, S.: The story of phosphorus: Global food security and food for thought. Global Environ. Chang. **19**, 292–305 (2009)
279. Cormack, R.S.: Discussion of "Tests of significance in 2 × 2 tables" by F. Yates. J. R. Stat. Soc. A Gen. **147**, 455 (1984)
280. Cormack, R.S.: The meaning of probability in relation to Fisher's exact test. Metron **44**, 1–30 (1986)
281. Cormack, R.S., Mantel, N.: Fisher's exact test: The marginal totals as seen from two different angles. Statistician **40**, 27–34 (1991)

282. Cornell alumni news. http://ecommons.cornell.edu/handle/1813/26700 (November 1923). Accessed 9 Sept 2011
283. Cornell alumni news. http://ecommons.cornell.edu/handle/1813/26818 (November 1926). Accessed 9 Sept 2011
284. Cornell alumni news. http://ecommons.cornell.edu/handle/1813/26498 (October 1918). Accessed 9 Sept 2011
285. Cornell alumni news. http://ecommons.cornell.edu/handle/1813/27006 (October 1931). Accessed 9 Sept 2011
286. Corsten, L.C.A., de Kroon, J.P.M.: Comment on "In dispraise of the exact test" by J. Berkson. J. Stat. Plan. Infer. **3**, 193–197 (1979)
287. Cottle, R.W.: George B. Dantzig: Operations research icon. Oper. Res. **53**, 892–898 (2005)
288. Cotton, W.R., Thompson, G., Mielke, P.W.: Realtime mesoscale prediction on workstations. Bull. Am. Meteorol. Soc. **75**, 349–362 (1994)
289. Coven, V.: A history of statistics in the social sciences. http://grad.usask.ca/gateway/art_Coven_spr_03.pdf (Spring 2003). Accessed 24 May 2012
290. Coveyou, R.R., Sullivan, J.G.: Algorithm 71: Permutation. Commun. ACM **4**, 497 (1961)
291. Cowan, D.W., Thompson, H.J., Paulus, H.J., Mielke, P.W.: Bronchial asthma associated with air pollutants from grain industry. JAPCA **13**, 546–552 (1963)
292. Cowles, M., Davis, C.: On the origins of the .05 level of statistical significance. Am. Psychol. **37**, 553–558 (1982)
293. Cox, D.F., Kempthorne, O.: Randomization tests for comparing survival curves. Biometrics **19**, 307–317 (1963)
294. Cox, D.R. (ed.): The Collected Works of John W. Tukey: Factorial and ANOVA, 1949–1962, vol. VII. Wadsworth Statistics and Probability Series. Wadsworth & Brooks/Cole, Pacific Grove (1992)
295. Cox, G.M., Homeyer, P.G.: Professional and personal glimpses of George W. Snedecor. Biometrics **31**, 265–301 (1975)
296. Craig, C.C.: The first course in mathematical statistics. Am. Stat. **5**, 14–16, 25 (1961)
297. Crow, J.F.: R. A. Fisher, a centennial view. Genetics **124**, 207–211 (1990)
298. Crowcroft, P.: Elton's Ecologists: A History of the Bureau of Animal Population. University of Chicago Press, Chicago (1991)
299. Cryan, M., Dyer, M.: A polynomial-time algorithm to approximately count contingency tables when the number of rows is constant. J. Comput. Syst. Sci. **67**, 291–310 (2003)
300. Curran-Everett, D.: Explorations in statistics: Standard deviations and standard errors. Adv. Physiol. Educ. **32**, 203–208 (2008)
301. Curran-Everett, D.: Explorations in statistics: Confidence intervals. Adv. Physiol. Educ. **33**, 87–90 (2009)
302. Curran-Everett, D.: Explorations in statistics: Hypothesis tests and P values. Adv. Physiol. Educ. **33**, 81–86 (2009)
303. Curran-Everett, D.: Explorations in statistics: The bootstrap. Adv. Physiol. Educ. **33**, 286–292 (2009)
304. Curran-Everett, D.: Explorations in statistics: Correlation. Adv. Physiol. Educ. **34**, 186–191 (2010)
305. Curran-Everett, D.: Explorations in statistics: Power. Adv. Physiol. Educ. **34**, 41–43 (2010)
306. Curran-Everett, D.: Explorations in statistics: Regression. Adv. Physiol. Educ. **35**, 347–352 (2011)
307. Curran-Everett, D.: Explorations in statistics: Permutation methods. Adv. Physiol. Educ. **36**, 181–187 (2012)
308. da Cruz, F.: Herman Hollerith. http://www.columbia.edu/cu/computinghistory/hollerith.html (2011). Accessed 12 Mar 2012
309. Dabrowska, D.M., Speed, T.P.: On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat. Sci. **5**, 465–480 (1990) [This is a translation by Dabrowska and Speed of an article originally published in Polish in *Annals of Agricultural Sciences* by Jerzy Spława-Neyman in 1923]

310. Dale, A.I.: A History of Inverse Probability from Thomas Bayes to Karl Pearson. Springer, New York (1991)
311. Dallal, G.E.: PITMAN: A FORTRAN program for exact randomization tests. Comput. Biomed. Res. **21**, 9–15 (1988)
312. Daniel, C.: Locating outliers in factorial experiments. Technometrics **2**, 149–156 (1960)
313. Daniel, W.W.: Applied Nonparametric Statistics. Houghton Mifflin, Boston (1978)
314. Daniels, H.E.: Rank correlation and population models (with discussion). J. R. Stat. Soc. B Met. **12**, 171–191 (1950)
315. Dantzig, G.B.: On the non-existence of tests of "Student's" hypothesis having power functions independent of σ. Ann. Math. Stat. **11**, 186–192 (1940)
316. Dantzig, G.B., Wald, A.: On the fundamental lemma of Neyman and Pearson. Ann. Math. Stat. **22**, 87–93 (1951)
317. Darnell, A.C.: Harold Hotelling 1896–1973. Stat. Sci. **3**, 57–62 (1988)
318. Darwin, C.: The Effects of Cross and Self Fertilization in the Vegetable Kingdom. John Murray, London (1876)
319. David, F.N.: Review of "Rank Correlation Methods" by M. G. Kendall. Biometrika **37**, 190 (1950)
320. David, F.N.: Games, Gods, and Gambling: The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era. Hafner, New York (1962)
321. David, H.A.: H.O. Hartley: 1912–1980. Int. Stat. Rev. **50**, 327–330 (1982)
322. David, H.A.: The Iowa State Statistical Laboratory: Antecedents and early years. In: David, H.A., David, H.T. (eds.) Statistics: An Appraisal, pp. 3–18. Iowa State University Press, Ames (1984)
323. David, H.A.: First (?) occurrence of common terms in mathematical statistics. Am. Stat. **49**, 121–133 (1995)
324. David, H.A.: Statistics in U.S. universities in 1933 and the establishment of the Statistical Laboratory at Iowa State. Stat. Sci. **13**, 66–74 (1998)
325. David, H.A.: Samuel Stanley Wilks (1906–1964). Am. Stat. **60**, 46–49 (2006)
326. David, H.A.: The beginnings of randomization tests. Am. Stat. **62**, 70–72 (2008)
327. David, H.A., David, H.T. (eds.): Statistics: An Appraisal. Iowa State University Press, Ames (1984)
328. David, S.T., Kendall, M.G., Stuart, A.: Some questions of distribution in the theory of rank correlation. Biometrika **38**, 131–140 (1951)
329. Davis, A.W., Speed, T.P.: An Edgeworth expansion for the distribution of the F-ratio under a randomization model for the randomized block design. In: Gupta, S.S., Berger, J.O. (eds.) Statistical Decision Theory and Related Topics IV, vol. II, pp. 119–130. Springer, New York (1988)
330. Davis, C.S.: A new approximation to the distribution of Pearson's chi-square. Stat. Sinica **3**, 189–196 (1993)
331. Death of Sir John Bennet Lawes. http://en.wikisource.org/wiki/The_Times/1900/Obituary/John_Bennet_Lawes (1 September 1900). Accessed 20 July 2012
332. Defining a legend. University of Wisconsin Foundation Insights. http://www.supportuw.org/wp-content/uploads/insights_04_fall.pdf (Fall 2004). Accessed 20 June 2013
333. de Mast, J.: Agreement and kappa-type indices. Am. Stat. **61**, 148–153 (2007)
334. de Montmort, P.R.: Essay d'Analyse sur les Jeux de Hazard (Analytical Essay on Gambling). Quillau, Paris (1708)
335. de Wet, T.: Statistics in the fifties. South African Statistical Association (2003) [Presidential Address on the 50th Anniversary of the South African Statistical Association, Johannesburg, 2003 by Tertius de Wet]
336. DeGroot, M.H.: A conversation with Charles Stein. Stat. Sci. **1**, 454–462 (1986)
337. DeGroot, M.H.: A conversation with Erich L. Lehmann. Stat. Sci. **1**, 243–258 (1986)
338. DeGroot, M.H.: A conversation with George Box. Stat. Sci. **2**, 239–258 (1987)
339. DeGroot, M.H.: A conversation with George A. Bernard. Stat. Sci. **3**, 196–212 (1988)

340. deLaubenfels, R.: The victory of least squares and orthogonality in statistics. Am. Stat. **60**, 315–321 (2006)
341. Denholm, C.: The history of Rothamsted research. Rothamsted Research. http://www.rothamsted.ac.uk/index.php (2012). Accessed 19 July 2012
342. Denker, M., Puri, M.L.: Asymptotic behavior of multi-response permutation procedures. Adv. Appl. Math. **9**, 200–210 (1988)
343. Denning, P.J.: The great principles of computing. Am. Sci. **98**, 369–372 (2010)
344. Dershowitz, N.: A simplified loop-free algorithm for generating permutations. BIT **15**, 158–164 (1975)
345. Deuchler, G.: Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie. Z. Padagog. Psychol. Exp. Padagog. **15**, 114–131, 145–159, and 229–242 (1914)
346. Diaconis, P., Freedman, D.: Finite exchangeable sequences. Ann. Probab. **8**, 745–764 (1980)
347. Diaconis, P., Graham, R.L.: Spearman's footrule as a measure of disarray. J. R. Stat. Soc. B Met. **39**, 262–268 (1977)
348. Diaconis, P., Holmes, S.: Gray codes for randomization procedures. Stat. Comput. **4**, 287–302 (1994)
349. Dielman, T.E.: A comparison of forecasts from least absolute and least squares regression. J. Forecasting **5**, 189–195 (1986)
350. Dielman, T.E.: Corrections to a comparison of forecasts from least absolute and least squares regression. J. Forecasting **8**, 419–420 (1989)
351. Dinneen, L.C., Blakesley, B.C.: Algorithm 62: A generator for the sampling distribution of the Mann–Whitney U statistic. J. R. Stat. Soc. C Appl. Stat. **22**, 269–273 (1973)
352. Dinneen, L.C., Blakesley, B.C.: Letter to the editors: Definition of Spearman's footrule. J. R. Stat. Soc. C Appl. Stat. **31**, 66 (1982)
353. Dixon, W.J.: A criterion for testing the hypothesis that two samples are from the same population. Ann. Math. Stat. **11**, 199–204 (1940)
354. Dodge, Y.: An introduction to statistical data analysis L_1-norm based. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-norm and Related Methods, pp. 1–21. Elsevier, Amsterdam (1987) [Collection of invited papers presented at The First International Conference on Statistical Data Analysis Based on the L_1-norm and Related Methods, held in Neuchâtel, Switzerland, from 31 August to 4 September 1987]
355. Dodge, Y.: A natural random number generator. Int. Stat. Rev. **64**, 329–344 (1996)
356. Dodge, Y. (ed.): The Oxford Dictionary of Statistical Terms. Oxford University Press, Oxford (2003)
357. Dolnick, E.: The Clockwork Universe: Isaac Newton, the Royal Society & the Birth of the Modern World. HarperCollins, New York (2011)
358. Donegani, M.: Asymptotic and approximate distribution of a statistic by resampling with or without replacement. Stat. Prob. Lett. **11**, 181–183 (1991)
359. Doolen, G.D., Hendricks, J.: Monte Carlo at work. Los Alamos Sci. **15**, 142–143 (1987)
360. Draper, N.R., Stoneman, D.M.: Testing for the inclusion of variables in linear regression by a randomization technique. Technometrics **8**, 695–699 (1966)
361. Dudycha, A.L., Dudycha, L.W.: Behavioral statistics: An historical perspective. In: Kirk, R.E. (ed.) Statistical Issues: A Reader for the Behavioral Sciences, pp. 2–25. Brooks/Cole, Belmont (1972)
362. Dukes, W.F.: $N = 1$. Psychol. Bull. **64**, 74–79 (1965)
363. Dunnett, C.W.: Frank Wilcoxon, 1892–1965. Technometrics **8**, 195–196 (1966)
364. Dupont, W.D.: Sensitivity of Fisher's exact test to minor perturbations in 2 × 2 contingency tables. Stat. Med. **5**, 629–635 (1986)
365. Dupont, W.D.: Reply to Comment on "Sensitivity of Fisher's exact test to minor perturbations in 2 × 2 contingency tables" by A. Martín Andrés and J.D. Luna del Castillo. Stat. Med. **8**, 244–245 (1989)
366. Duran, B.S., Mielke, P.W.: Robustness of sum of squared ranks test. J. Am. Stat. Assoc. **63**, 338–344 (1968)
367. Durstenfeld, R.: Algorithm 235: Random permutation. Commun. ACM **7**, 420 (1964)

368. Dwass, M.: Modified randomization tests for nonparametric hypotheses. Ann. Math. Stat. **28**, 181–187 (1957)
369. Dyke, G.: Obituary: Frank Yates. J. R. Stat. Soc. A Stat. **158**, 333–338 (1995)
370. Dyson, G.: Turing's Cathedral: The Origins of the Digital Universe. Pantheon/Vintage, New York (2012)
371. Eaves, B.C.: Algorithm 130: PERMUTE. Commun. ACM **5**, 551 (1962)
372. Eckhardt, R.: Stan Ulam, John von Neumann, and the Monte Carlo method. Los Alamos Sci. **15**, 131–137 (1987)
373. Eddy, W.F., Huber, P.J., McClure, D.E., Moore, D.S., Stuetzle, W., Thisted, R.A.: Computers in statistical research. Stat. Sci. **1**, 419–437 (1986)
374. Eden, T.: Soil Erosion. Imperial Bureau of Soil Science, Rothamsted Experimental Station (1933)
375. Eden, T.: Elements of Tropical Soil Science. MacMillan, London (1947)
376. Eden, T.: Tea. Tropical Agriculture. Longmans, Green, London (1958)
377. Eden, T., Fisher, R.A.: Studies in crop variation, IV. The experimental determination of the value of top dressings with cereals. J. Agric. Sci. **17**, 548–562 (1927)
378. Eden, T., Fisher, R.A.: Studies in crop variation, VI. Experiments on the response of the potato to potash and nitrogen. J. Agric. Sci. **19**, 201–213 (1929)
379. Eden, T., Yates, F.: On the validity of Fisher's z test when applied to an actual example of non-normal data. J. Agric. Sci. **23**, 6–17 (1933)
380. Edgeworth, F.Y.: The method of least squares. Philos. Mag. 5 **16**, 360–375 (1883)
381. Edgeworth, F.Y.: The choice of means. Philos. Mag. 5 **24**, 268–271 (1887)
382. Edgeworth, F.Y.: Letter calling attention to article in *Hermathena*. Philos. Mag. 5 **24**, 222–223 (1887)
383. Edgeworth, F.Y.: On a new method of reducing observations relating to several quantities. Philos. Mag. 5 **25**, 184–191 (1887)
384. Edgeworth, F.Y.: On discordant observations. Lond. Edin. Dub. Philos. Mag. **23**, 364–375 (1887)
385. Edgeworth, F.Y.: On observations relating to several quantities. Hermathena **6**, 279–285 (1887)
386. Edgeworth, F.Y.: On the use of medians for reducing observations relating to several quantities. Philos. Mag. 6 **46**, 1074–1088 (1923)
387. Edgington, E.S.: Randomization tests. J. Psychol. **57**, 445–449 (1964)
388. Edgington, E.S.: Statistical inference and nonrandom samples. Psychol. Bull. **66**, 485–487 (1966)
389. Edgington, E.S.: Statistical inference from $N = 1$ experiments. J. Psychol. **65**, 195–199 (1967)
390. Edgington, E.S.: Approximate randomization tests. J. Psychol. **72**, 143–149 (1969)
391. Edgington, E.S.: Statistical Inference: The Distribution-free Approach. McGraw-Hill, New York (1969)
392. Edgington, E.S.: Randomization Tests. Marcel Dekker, New York (1980)
393. Edgington, E.S.: Randomization Tests, 2nd edn. Marcel Dekker, New York (1987)
394. Edgington, E.S.: Randomization Tests, 3rd edn. Marcel Dekker, New York (1995)
395. Edgington, E.S., Khuller, P.L.V.: A randomization test computer program for trends in repeated-measures data. Educ. Psychol. Meas. **52**, 93–95 (1992)
396. Edgington, E.S., Onghena, P.: Randomization Tests, 4th edn. Chapman & Hall/CRC, Boca Raton (2007)
397. Edgington, E.S., Strain, A.R.: Randomization tests: Computer time requirements. J. Psychol. **85**, 89–95 (1973)
398. Edwards, A.E.W., Bodmer, W.: R. A. Fisher — 50 years on. Significance **9**, 27–29 (December, 2012)
399. Edwards, A.W.F.: Pascal's Arithmetical Triangle. Griffin, London (1987)
400. Edwards, A.W.F.: Professor C. A. B. Smith, 1917–2002. J. R. Stat. Soc. D Stat. **51**, 404–405 (2002)

401. Edwards, A.W.F.: Fisher computes.... Significance **9**, 44 (2012) [Letter to the editor regarding an article by Fisher in *Significance*, August 2012]
402. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton (1993)
403. Ehrlich, G.: Algorithm 466: Four combinatorial algorithms. Commun. ACM **16**, 690–691 (1973)
404. Ehrlich, G.: Loopless algorithms for generating permutations, combinations, and other combinatorial configurations. J. ACM **20**, 500–513 (1973)
405. Eicker, P.J., Siddiqui, M.M., Mielke, P.W.: A matrix occupancy problem. Ann. Math. Stat. **43**, 988–996 (1972)
406. Churchill Eisenhart, statistics expert. NY Times. http://www.nytimes.com/1994/07/01/us/churchill-eisenhart-statistics-expert-82.html (1 July 1994). Accessed 20 Jan 2012
407. Eisenhart, C.: Boscovich and the combination of observations. In: Roger Joseph Boscovich S.J., F.R.S., 1711–1787. Studies of His Life and Work on the 250th Anniversary of His Birth, pp. 200–212. Allen & Unwin, London (1961)
408. Eisenhart, C.: On the transition from "Student's" z to "Student's" t. Am. Stat. **33**, 6–10 (1979)
409. Ellis, M.G., Heckstall-Smith, H.W.: Fun with statistics. Tubercle **36**, 378–381 (1955)
410. Endler, J.A., Mielke, P.W.: Comparing entire colour patterns as birds see them. Biol. J. Linn. Soc. **86**, 405–431 (2005)
411. Entsuah, A.R.: Randomization procedures for analyzing clinical trial data with treatment related withdrawals. Commun. Stat. Theor. M. **19**, 3859–3880 (1990)
412. Epstein, D.M., Dalinka, M.K., Kaplan, F.S., Aronchick, J.M., Marinelli, D.L., Kundel, H.L.: Observer variation in the detection of osteopenia. Skeletal Radiol. **15**, 347–349 (1986)
413. Ernst, M.D.: Permutation methods: A basis for exact inference. Stat. Sci. **19**, 676–685 (2004)
414. Esscher, F.: On a method of determining correlation from the ranks of variates. Skand. Aktuarietidskr. **7**, 201–219 (1924)
415. Everitt, B.S.: Moments of the statistics kappa and weighted kappa. Br. J. Math. Stat. Psychol. **21**, 97–103 (1968)
416. Faith, D.P.: Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. **40**, 366–375 (1991)
417. Faith, D.P., Cranston, P.S.: Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. Cladistics **7**, 1–28 (1991)
418. Faith, D.P., Trueman, J.W.H.: When the topology-dependent permutation test (T-PTP) for monophyly returns significant support for monophyly, should that be equated with (a) rejecting a null hypothesis of nonmonophyly, (b) rejecting a null hypothesis of "no structure," (c) failing to falsify a hypothesis of monophyly, or (d) none of the above? Syst. Biol. **45**, 580–586 (1996)
419. Farebrother, R.W.: The historical development of the L_1 and L_∞ estimation procedures, 1793–1930. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-Norm and Related Methods, pp. 37–63. North-Holland, Amsterdam (1987)
420. Farebrother, R.W.: Rogerius Josephus Boscovich. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 82–85. Springer, New York (2001)
421. Feinstein, A.R.: Clinical Biostatistics XXIII: The role of randomization in sampling, testing, allocation, and credulous idolatry (Part 2). Clin. Pharmacol. Ther. **14**, 898–915 (1973)
422. Feinstein, A.R.: Clinical Biostatistics. C. V. Mosby, St. Louis (1977)
423. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol. **43**, 543–549 (1990)
424. Feldman, S.E., Klinger, E.: Shortcut exact calculation of the Fisher–Yates "exact test". Psychometrika **28**, 289–291 (1963)
425. Fernholz, L.T., Morganthaler, S., Tukey, J.W., Tukey, E.: A conversation with John W. Tukey and Elizabeth Tukey. Stat. Sci. **15**, 79–95 (2000)
426. Ferry, G.: A Computer Called LEO. HarperCollins, London (2004)

427. Festinger, L.: The significance of differences between means without reference to the frequency distribution function. Psychometrika **11**, 97–105 (1946)
428. Fienberg, S.E.: A brief history of statistics in three and one-half chapters: A review essay. Hist. Method. **24**, 124–135 (1991)
429. Fienberg, S.E., Stigler, S.M., Tanur, J.M.: The William Kruskal legacy: 1919–2005. Stat. Sci. **22**, 255–261 (2007)
430. Fienberg, S.E., Tanur, J.M.: Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. Int. Stat. Rev. **64**, 237–253 (1966)
431. Fienberg, S.E., Tanur, J.M.: Jerzy Neyman. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 444–448. Springer, New York (2001)
432. Fike, C.T.: A permutation generation method. Comput. J. **18**, 21–22 (1975)
433. Finch, W.H., Davenport, T.: Performance of Monte Carlo permutation and approximate tests for multivariate means comparisons with small sample sizes when parametric assumptions are violated. Methodology **5**, 60–70 (2009)
434. Finney, D.J.: The Fisher–Yates test of significance in 2 × 2 contingency tables. Biometrika **35**, 145–156 (1948)
435. Finney, D.J.: A numerate life. In: Gani, J. (ed.) The Making of Statisticians, pp. 150–164. Springer, New York (1982)
436. Finney, D.J.: Frank Yates: 12 May 1902 — 17 June 1994. Bio. Mem. Fellows Roy. Soc. **41**, 554–573 (1995). [Until 1955, Biographical Memoirs of Fellows of the Royal Society was known as Obituary Notices of Fellows of the Royal Society.]
437. Finney, D.J.: Remember a pioneer: Frank Yates (1902–1994). Teach. Stat. **20**, 2–5 (1998)
438. Finney, D.J.: Calibration guidelines challenge outlier practices. Am. Stat. **60**, 309–314 (2006)
439. Finney, D.J., Latscha, R., Bennett, B., Hsu, P.: Tables for Testing Significance in a 2 × 2 Contingency Table. Cambridge University Press, London (1963)
440. Fisher, R.A.: Untitled. Nature **124**, 266–267 (17 August 1929)
441. Fisher, R.A.: On an absolute criterion for fitting frequency curves. Mess. Math. **41**, 155–160 (1912)
442. Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika **10**, 507–521 (1915)
443. Fisher, R.A.: The causes of human variability. Eugen. Rev. **10**, 213–220 (1918)
444. Fisher, R.A.: On the "probable error" of a coefficient of correlation deduced from a small sample. Metron **1**, 1–32 (1921)
445. Fisher, R.A.: Studies in crop variation, I. An examination of the yield of dressed grain from Broadbalk. J. Agric. Sci. **11**, 107–135 (1921)
446. Fisher, R.A.: On the interpretation of χ^2 from contingency tables, and the calculation of p. J. R. Stat. Soc. **85**, 87–94 (1922)
447. Fisher, R.A.: The distribution of the partial correlation coefficient. Metron **3**, 329–332 (1924)
448. Fisher, R.A.: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh (1925)
449. Fisher, R.A.: The arrangement of field experiments. J. Am. Stat. Assoc. **33**, 503–513 (1926)
450. Fisher, R.A.: Statistical Methods for Research Workers, 5th edn. Oliver and Boyd, Edinburgh (1934)
451. Fisher, R.A.: The Design of Experiments. Oliver and Boyd, Edinburgh (1935)
452. Fisher, R.A.: The logic of inductive inference (with discussion). J. R. Stat. Soc. **98**, 39–82 (1935)
453. Fisher, R.A.: 'The coefficient of racial likeness' and the future of craniometry. J. R. Anthropol. Inst. **66**, 57–63 (1936)
454. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenic. **7**, 179–188 (1936)
455. Fisher, R.A.: Statistical Theory of Estimation. University of Calcutta Press, Calcutta (1938)
456. Fisher, R.A.: The interpretation of experimental four-fold tables. Science **94**, 210–211 (1941)
457. Fisher, R.A.: A new test for 2 × 2 tables. Nature **156**, 388 (1945)

458. Fisher, R.A.: Gene frequencies in a cline determined by selection and diffusion. Biometrics **6**, 353–361 (1950)
459. Fisher, R.A.: Mathematics of a lady tasting tea. In: Newman, J.R. (ed.) The World of Mathematics, vol. III, section VIII, pp. 1512–1521. Simon & Schuster, New York (1956)
460. Fisher, R.A.: Statistical Methods and Scientific Inference, 2nd edn. Hafner, New York (1959)
461. Fisher, R.A.: The Design of Experiments, 7th edn. Hafner, New York (1960)
462. Fisher, R.A., Mackenzie, W.A.: Studies in crop variation, II. The manurial response of different potato varieties. J. Agric. Sci. **13**, 311–320 (1923)
463. Fisher, R.A., Yates, F.: Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, Edinburgh (1938)
464. Fitzmaurice, G.M., Lipsitz, S.R., Ibrahim, J.G.: A note on permutation tests for variance components in multilevel generalized linear mixed models. Biometrics **63**, 942–946 (2007)
465. Fix, E., Hodges, J.L.: Significance probabilities of the Wilcoxon test. Ann. Math. Stat. **26**, 301–312 (1955)
466. Fleishman, A.I.: A program for calculating the exact probability along with explorations of M by N contingency tables. Educ. Psychol. Meas. **37**, 799–803 (1977)
467. Fleiss, J.L.: Statistical Methods for Rates and Proportions, 2nd edn. Wiley, New York (1981)
468. Fleiss, J.L., Cicchetti, D.V.: Inference about weighted kappa in the non-null case. Appl. Psychol. Meas. **2**, 113–117 (1978)
469. Fleiss, J.L., Cohen, J., Everitt, B.S.: Large sample standard errors of kappa and weighted kappa. Psychol. Bull. **72**, 323–327 (1969)
470. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions, 5th edn. Wiley, New York (2003)
471. Fligner, M.A.: Comment on “Rank transformations as a bridge between parametric and nonparametric statistics” by W.J. Conover and R.L. Iman. Am. Stat. **35**, 131–132 (1981)
472. Flournoy, N.: A conversation with Wilfrid J. Dixon. Stat. Sci. **8**, 458–477 (1993)
473. Flournoy, N.: Wilfrid Joseph Dixon, 1915–2008. J. R. Stat. Soc. A Stat. **173**, 455–456 (2010)
474. Folks, J.L.: A conversation with Oscar Kempthorne. Stat. Sci. **10**, 321–336 (1995)
475. Følling, A.: The excretion of phenylpyruvic acid in the urine, an anomaly of metabolism in connection with imbecility. Z. Physiol. Chem. **227**, 169–176 (1934)
476. Fraker, M.E., Peacor, S.D.: Statistical tests for biological interactions: A comparison of permutation tests and analysis of variance. Acta Oecol. **33**, 66–72 (2008)
477. Franklin, L.A.: Exact tables of Spearman’s footrule for $n = 11(1)18$ with estimate of convergence and errors for the normal approximation. Stat. Prob. Lett. **6**, 399–406 (1988)
478. Freedman, D., Lane, D.: A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. **1**, 292–298 (1983)
479. Freedman, D.A., Pisani, R., Purves, R., Adhikari, A.: Statistics, 2nd edn. Norton, New York (1991)
480. Freeman, G.H., Halton, J.H.: Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika **38**, 141–149 (1951)
481. Freund, J.E., Ansari, A.R.: Two-way rank sum test for variances. Tech. Rep. 34, Virginia Polytechnic and State University, Blacksburg (1957)
482. Frick, R.W.: Interpreting statistical testing: Process and propensity, not population and random sampling. Behav. Res. Methods Instrum. C **30**, 527–535 (1998)
483. Milton Friedman’s Bio. The Milton Friedman Foundation for Educational Choice. http://www.edchoice.org/The-Friedmans/Milton-Friedman-s-Bio.aspx. Accessed 20 Dec 2011
484. Friedman, L.M., Furberg, C.D., DeMets, D.L.: Fundamentals of Clinical Trials, 3rd edn. Springer, New York (1998)
485. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. **32**, 675–701 (1937)
486. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Ann. Math. Stat. **11**, 86–92 (1940)
487. Friedman, M.: Autobiography. The Official Website of the Nobel Prize. http://www.nobelprize.org/nobel_prizes/economics/laureates/1976/friedman-autobio.html (2005). Accessed 20 Dec 2011

488. Fuechsle, M., Miwa, J.A., Mahapatra, S., Ryu, H., Lee, S., Warschkow, O., Hollenberg, L.C.L., Klimeck, G., Simmons, M.Y.: A single-atom transistor. Nat. Nanotechnol. http://www.nature.com/nnano/journal/vaop/ncurrent/full/nnano.2012.21.html (19 February 2012). Accessed 25 Feb 2012
489. Gabriel, K.R., Hall, W.J.: Rerandomization inference on regression and shift effects: Computationally feasible methods. J. Am. Stat. Assoc. **78**, 827–836 (1983)
490. Gail, M., Mantel, N.: Counting the number of $r \times c$ contingency tables with fixed margins. J. Am. Stat. Assoc. **72**, 859–862 (1977)
491. Gail, M.H., Tan, W.Y., Piantadosi, S.: Tests for no treatment effect in randomized clinical trials. Biometrika **75**, 57–64 (1988)
492. Galilei, G.: Dialogo sopra i due massimi sistemi del mondo: Tolemaico, e Copernicano. Landini, Florence (1632). [English translation, *Dialogue Concerning the Two Chief World Systems, Ptolemaic and Copernican*, by Stillman Drake (with foreword by Albert Einstein), published by the University of California Press, Berkeley, CA, 1953.]
493. Gani, J. (ed.): The Making of Statisticians. Springer, New York (1982)
494. Gans, L.P., Robertson, C.A.: Distributions of Goodman and Kruskal's gamma and Spearman's rho in 2×2 tables for small and moderate sample sizes. J. Am. Stat. Assoc. **76**, 942–946 (1981)
495. Garfield, S.: Just My Type: A Book About Fonts. Gotham Books, New York (2011)
496. Bill Gates. Biography.com. http://www.biography.com/people/bill-gates-9307520 (2012). Accessed 20 Oct 2012
497. Gates, B.: The Road Ahead. Penguin, New York (1995)
498. Gates, B.: Business @ the Speed of Thought. Penguin, New York (1999)
499. Gayen, A.K.: The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika **38**, 219–247 (1951)
500. Geary, R.C.: Some properties of correlation and regression in a limited universe. Metron **7**, 83–119 (1927)
501. Geary, R.C.: Testing for normality. Biometrika **34**, 209–242 (1947)
502. Gebhard, J., Schmitz, N.: Permutation tests — a revival?! I. Optimum properties. Stat. Pap. **39**, 75–85 (1998)
503. Gebhard, J., Schmitz, N.: Permutation tests — a revival?! II. An efficient algorithm for computing the critical region. Stat. Pap. **39**, 87–96 (1998)
504. Geissler, A.: Beitrage sur frag des geschlechtsverkältnisses der geborenen. . Konig. Sach. Stat. Bur. **35**, 1–24 (1889)
505. Gelman, A.: Tables as graphs: The Ramanujan principle. Significance **8**, 183 (December 2011)
506. Gentle, J.E.: Least absolute values estimation: An introduction. Commun. Stat. Simul. C **6**, 313–328 (1977)
507. Gentleman, J.F.: Generation of all ${}_NC_R$ combinations by simulating nested FORTRAN *DO* loops. J. R. Stat. Soc. C Appl. **24**, 374–376 (1975)
508. Gerhardt, C.I. (ed.): Die Mathematische Schriften von Gottfried Wilhelm Leibniz, vol. 7. Weidmann, Berlin (1855)
509. Ghent, A.W.: Inside Illinois, University of Illinois at Urbana-Champaign. http://news.illinois.edu/ii/01/0503/0503index.html#deaths (2 May 2001). Accessed 28 Mar 2012
510. Ghent, A.W.: A method for exact testing of 2×2, 2×3, 3×3, and other contingency tables, employing binomial coefficients. Am. Midl. Nat. **88**, 15–27 (1972)
511. Gibbons, J.D., Pratt, J.W.: P-values: Interpretation and methodology. Am. Stat. **29**, 20–25 (1975)
512. Gigerenzer, G., Swijtink, Z., Porter, T.M., Daston, L.: The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge University Press, Cambridge (1989)
513. Gill, P.M.W.: Efficient calculation of p-values in linear-statistic permutation significance tests. J. Stat. Comput. Simul. **77**, 55–61 (2007)

514. Gittelsohn, A.M.: An occupancy problem. Am. Stat. **23**, 11–12 (1969)
515. Gladwell, M.: Outliers: The Story of Success. Little, Brown, New York (2008)
516. Glasser, G.J., Winter, R.F.: Critical values of the coefficient of rank correlation for testing the hypothesis of independence. Biometrika **48**, 444–448 (1961)
517. Gluud, C.: Trials and errors in clinical research. Lancet **354**, SIV59 (1999)
518. Good, I.J.: On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. Ann. Stat. **4**, 1159–1189 (1976)
519. Good, I.J.: The early history of the Fisher–Yates–Irwin formula and Fisher's 'exact test'. J. Stat. Comput. Simul. **19**, 315–319 (1984)
520. Good, I.J.: A further note on the history of the Fisher–Yates–Irwin formula. J. Stat. Comput. Simul. **20**, 155–159 (1984)
521. Good, I.J.: Further comments concerning the lady tasting tea or beer: P-values and restricted randomization. J. Stat. Comput. Simul. **40**, 263–267 (1992)
522. Good, P.I.: Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer, New York (1994)
523. Good, P.I.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer, New York (1994)
524. Good, P.I.: Resampling Methods: A Practical Guide to Data Analysis. Birkhäuser, Boston (1999)
525. Good, P.I.: Permutation, Parametric and Bootstrap Tests of Hypotheses, 2nd edn. Springer, New York (2000)
526. Good, P.I.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edn. Springer, New York (2000)
527. Good, P.I.: Resampling Methods: A Practical Guide to Data Analysis, 2nd edn. Birkhäuser, Boston (2001)
528. Good, P.I.: Extensions of the concept of exchangeability and their applications. J. Mod. Appl. Stat. Methods **1**, 243–247 (2002)
529. Good, P.I.: Efficiency comparisons of rank and permutation tests by Janice M. Weinberg and Stephen W. Lagakos in *Statistics in Medicine* 2001; **20**:705–731. Stat. Med. **23**, 857 (2004)
530. Good, P.I.: Efficiency comparisons of rank and permutation tests by Phillip I. Good in *Statistics in Medicine* 2004; **23**:857. Stat. Med. **24**, 1777–1781 (2005)
531. Good, P.I.: Permutation, Parametric and Bootstrap Tests of Hypotheses, 3rd edn. Springer, New York (2005)
532. Good, P.I.: Resampling Methods: A Practical Guide to Data Analysis, 3rd edn. Birkhäuser, Boston (2006)
533. Good, P.I., Xie, F.: Analysis of a crossover clinical trial by permutation methods. Contemp. Clin. Trials **29**, 565–568 (2008)
534. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. J. Am. Stat. Assoc. **49**, 732–764 (1954)
535. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, II: Further discussion and references. J. Am. Stat. Assoc. **54**, 123–163 (1959)
536. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, III: Approximate sampling theory. J. Am. Stat. Assoc. **58**, 310–364 (1963)
537. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, IV: Simplification of asymptotic variances. J. Am. Stat. Assoc. **67**, 415–421 (1972)
538. Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross Classifications. Springer, New York (1979)
539. Goodman, S.N.: p values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate (with discussion). Am. J. Epidemiol. **137**, 485–501 (1993)
540. Google Company Management Team. http://www.google.com/about/company/facts/management/ (2012). Accessed 3 Nov 2012
541. Gordon, A.D., Buckland, S.T.: A permutation test for assessing the similarity of ordered sequences. Math. Geol. **28**, 735–742 (1996)

542. Gover, J.E.B., Mawer, A., Stenton, F.M.: The Place-names of Hertfordshire, vol. 15. Cambridge University Press, Cambridge (1938) [Published by the English Place-name Society]
543. Graham, P.: Author's reply to "Comment on: Modeling covariate effects in observer agreement studies: The case of nominal scale agreement" by I. Guggenmoos-Holzmann. Stat. Med. **14**, 2286–2288 (1995)
544. Graham, P., Jackson, R.: The analysis of ordinal agreement data: Beyond weighted kappa. J. Clin. Epidemiol. **46**, 1055–1062 (1993)
545. Grant, L.O., Mielke, P.W.: A randomized cloud seeding experiment at Climax, Colorado, 1960–65. In: Cam, L.M., Neyman, J. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. V, pp. 115–131. University of California Press, Berkeley (1967)
546. Graves, T., Reese, C.S., Fitzgerald, M.: Hierarchical models for permutations: Analysis of auto racing results. J. Am. Stat. Assoc. **98**, 282–291 (2003)
547. Gray, W.M., Landsea, C.W., Mielke, P.W., Berry, K.J.: Predicting Atlantic seasonal hurricane activity 6–11 months in advance. Weather Forecast. **7**, 440–455 (1992)
548. Green, B.F.: A practical interactive program for randomization tests of location. Am. Stat. **31**, 37–39 (1977)
549. Green, B.F.: Randomization tests. J. Am. Stat. Assoc. **76**, 495 (1981) [Review of E.S. Edgington's *Randomization Tests* by Bert F. Green]
550. Greenberg, B.G.: Joseph Oscar Irwin 1898–1982. Biometrics **39**, 527–528 (1983)
551. Greenland, S.: On the logical justification of conditional tests for two-by-two contingency tables. Am. Stat. **45**, 248–251 (1991)
552. Greenwood, M.: The statistical study of infectious diseases (with discussion). J. R. Stat. Soc. **109**, 85–110 (1946)
553. Gregory, R.J.: A FORTRAN computer program for the Fisher exact probability test. Educ. Psychol. Meas. **33**, 697–700 (1973)
554. Greiner, R.: Über das Fehlersystem der Kollektivmasslehre. Z. Math. Phys. **57**, 121, 225, 337 (1909)
555. Gridgeman, N.T.: The lady tasting tea, and allied topics. J. Am. Stat. Assoc. **54**, 776–783 (1959)
556. Grier, D.A.: Statistics and the introduction of digital computers. Chance **4**, 30–36 (1991)
557. Grier, D.A.: Statistical laboratories and the origins of computing. Chance **12**, 14–20 (1999)
558. Griffin, H.D.: Graphic computation of tau as a coefficient of disarray. J. Am. Stat. Assoc. **53**, 441–447 (1958)
559. Grosenbaugh, L.R.: More on FORTRAN random number generators. Commun. ACM **12**, 369 (1969)
560. Guggenmoos-Holzmann, I.: How reliable are chance-corrected measures of agreement? Stat. Med. **12**, 2191–2205 (1993)
561. Guggenmoos-Holzmann, I.: Comment on "Modeling covariate effects in observer agreement studies: The case of nominal scale agreement" by P. Graham. Stat. Med. **14**, 2285–2286 (1995)
562. Guide to the Meyer Dwass (1923–1996) papers. Northwestern University Library. http://findingaids.library.northwestern.edu/catalog/inu-ead-nua-archon-548 (2002). Accessed 19 Jan 2012
563. Gumbel, E.J.: Discussion of the papers of Messrs. Anscombe and Daniel. Technometrics **2**, 165–166 (1960)
564. Haber, M.: A comparison of some continuity corrections for the χ^2 test on 2×2 tables. J. Am. Stat. Assoc. **75**, 510–515 (1980)
565. Haber, M.: Comments on "The test of homogeneity for 2×2 contingency tables: A review of and some personal opinions on the controversy" by G. Camilli. Psychol. Bull. **108**, 146–149 (1990)
566. Hack, H.R.B.: An empirical investigation into the distribution of the F-ratio in samples from two non-normal populations. Biometrika **45**, 260–265 (1958)

567. Hacking, I.: The Emergence of Probability. Cambridge University Press, Cambridge (1975)
568. Hacking, I.: The Taming of Chance. Cambridge University Press, Cambridge (1990)
569. Haden, H.G.: A note on the distribution of the different orderings of *n* objects. Math. Proc. Cambridge **43**, 1–9 (1947)
570. Hald, A.: Statistical Tables and Formulas. Wiley, New York (1952)
571. Hald, A.: History of Probability and Statistics and Their Applications Before 1750. Wiley, New York (1990)
572. Hald, A.: A History of Mathematical Statistics from 1750 to 1930. Wiley, New York (1998)
573. Haldane, J.B.S., Smith, C.A.B.: A simple exact test for birth-order effect. Ann. Eugenic. **14**, 117–124 (1948)
574. Hall, A.D.: The Book of the Rothamsted Experiments. John Murray, London (1905)
575. Hall, N.S.: R. A. Fisher and his advocacy of randomization. J. Hist. Biol. **40**, 295–325 (2007)
576. Hall, N.S.: Ronald Fisher and Gertrude Cox: Two statistical pioneers sometimes cooperate and sometimes collide. Am. Stat. **64**, 212–220 (2010)
577. Hall, P., Wilson, S.R.: Two guidelines for bootstrap hypothesis testing. Biometrics **47**, 757–762 (1991)
578. Halton, J.H.: A rigorous derivation of the exact contingency formula. Math. Proc. Cambridge **65**, 527–530 (1969)
579. Halton, J.H.: A retrospective and prospective survey of the Monte Carlo method. SIAM Rev. **12**, 1–63 (1970)
580. Hamilton, A.: Brains that click. Pop. Mech. **91**, 162–167, 256, 258 (1949)
581. Hammersley, J.M.: J. Neyman, 1894–1981. J. R. Stat. Soc. A Gen. **145**, 523–524 (1982)
582. Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)
583. Hancock, T.W.: Remark on algorithm 434. Commun. ACM **18**, 117–119 (1975)
584. Hanley, J.A.: Standard error of the kappa statistic. Psychol. Bull. **102**, 315–321 (1987)
585. Hardy, G.H., Ramanujan, S.: Asymptotic formulae in combinatory analysis. Proc. Lond. Math. Soc. **17**, 75–115 (1918)
586. Harriot, T.: Thomas Harriot College of Arts and Sciences. http://www.ecu.edu/cs-cas/harriot/index.cfm (2009). Accessed 9 Sept 2012
587. Harriot, T.: A Briefe and True Report of the New Found Land of Virginia: of the Commodities and of the Nature and Manners of the Naturall Inhabitants : Discouered bÿ the English Colonÿ There Seated by Sir Richard Greinuile Knight In the ÿeere 1585 : Which Remained Vnder the Gouerenment of Twelue Monethes, At the Special Charge and Direction of the Honourable Sir Walter Raleigh Knight Lord Warden of the Stanneries Who therein Hath Beene Fauoured and Authorised bÿ Her Maiestie and Her Letters Patents / This Fore Booke Is Made in English by Thomas Hariot seruant to the Aboue-Named Sir Walter, a Member of the Colonÿ, and There Imploÿed in Discouering. University of North Carolina at Chapel Hill, Chapel Hill (2003). Electronic version available at http://docsouth.unc.edu/nc/hariot/hariot.html
588. Harris, T.E.: Regression using minimum absolute deviations. Am. Stat. **4**, 14–15 (1950)
589. Harter, H.L.: The method of least squares and some alternatives: Part I. Int. Stat. Rev. **42**, 147–174 (1974)
590. Harter, H.L.: The method of least squares and some alternatives: Part II. Int. Stat. Rev. **42**, 235–264, 282 (1974)
591. Harter, H.L.: The method of least squares and some alternatives: Addendum to Part IV. Int. Stat. Rev. **43**, 273–278 (1975)
592. Harter, H.L.: The method of least squares and some alternatives: Part III. Int. Stat. Rev. **43**, 1–44 (1975)
593. Harter, H.L.: The method of least squares and some alternatives: Part IV. Int. Stat. Rev. **43**, 125–190 (1975)
594. Harter, H.L.: The method of least squares and some alternatives: Part V. Int. Stat. Rev. **43**, 269–272 (1975)

595. Harter, H.L.: The method of least squares and some alternatives: Part VI. Subject and author indexes. Int. Stat. Rev. **44**, 113–159 (1976)
596. Hartley, H.O.: The use of range in analysis of variance. Biometrika **37**, 271–280 (1950)
597. Hatfield, M.O.: Henry Agard Wallace (1941–1945). In: Vice Presidents of the United States, 1789–1993, pp. 399–406. United States Government Printing Office, Washington, DC (1997)
598. Hauben, R.: History of the computer: Part 4. Amat. Comput. **2**, 9–12 (1989)
599. Haviland, M.G.: Yates's correction for continuity and the analysis of 2×2 contingency tables (with discussion). Stat. Med. **9**, 363–367 (1990)
600. Hayes, A.F.: Permustat: Randomization tests for the Macintosh. Behav. Res. Methods Instrum. C **28**, 473–475 (1996)
601. Hayes, A.F.: Permutation test is not distribution-free: Testing $H_0\colon \rho = 0$. Psychol. Method. **1**, 184–198 (1996)
602. Hayes, B.: The memristor. Am. Sci. **99**, 106–110 (2011)
603. Hays, J.E.: A FORTRAN procedure for Fisher's exact probability test. Behav. Res. Methods Instrum. **7**, 481 (1975)
604. Healy, M.J.R.: Discussion of "Tests of significance in 2×2 tables" by F. Yates. J. R. Stat. Soc. A Gen. **147**, 456–457 (1984)
605. Healy, M.J.R.: Frank Yates 1902–1994. Biometrics **51**, 389–391 (1995)
606. Healy, M.J.R.: Frank Yates, 1902–1994 — The work of a statistician. Int. Stat. Rev. **63**, 271–288 (1995)
607. Healy, M.J.R.: R. A. Fisher the statistician. Statistician **52**, 303–310 (2003)
608. Heap, B.R.: Permutations by interchanges. Comput. J. **6**, 293 (1963)
609. Hellman, M.: A study of some etiological factors of malocclusion. Dent. Cosmos **56**, 1017–1032 (1914)
610. Hemelrijk, J.: Note on Wilcoxon's two-sample test when ties are present. Ann. Math. Stat. **23**, 133–135 (1952)
611. Herbert, D.A.: Statistics in U.S. universities in 1933 and the establishment of the Statistical Laboratory at Iowa State. Stat. Sci. **13**, 66–74 (1998)
612. Herman, P.G., Khan, A., Kallman, C.E., Rojas, K.A., Carmody, D.P., Bodenheimer, M.M.: Limited correlation of left ventricular end-diastolic pressure with radiographic assessment of pulmonary hemodynamics. Radiology **174**, 721–724 (1990)
613. Heyde, C.C., Seneta, E. (eds.): Statisticians of the Centuries. Springer, New York (2001)
614. Higgins, J.J., Blair, R.C.: Comment on "Why permutation tests are superior to t and F tests in biomedical research" by J. Ludbrook and H.A.F. Dudley. Am. Stat. **54**, 86 (2000)
615. Higgon, K.: Rosenbaum, Dr Sidney (b 1918). Liddell Hart Centre for Military Archives, King's College London. http://www.kingscollections.org/catalogues/lhcma/collection/p-t/ro75-001 (2007). Accessed 6 Dec 2012
616. Higgs, M.D.: Do we really need the *S*-word? Am. Sci. **101**, 6–8. http://www.americanscientist.org/issues/pub/2013/1/do-we-really-need-the-s-word (2013). Accessed 4 Jan 2013
617. Hilbe, J.: Section editor's notes: Statistical computing software reviews. Am. Stat. **54**, 71 (2000)
618. Hilbe, J.M.: The coevolution of statistics and HZ. In: Sawilowsky, S.S. (ed.) Real Data Analysis, pp. 3–20. Information Age, Charlotte (2007)
619. Hilbert, M.: How much information is there in the "information society?". Significance **9**, 8–12 (2012)
620. Hill, I.D.: Discussion of "Tests of significance in 2×2 tables" by F. Yates. J. R. Stat. Soc. A Gen. **147**, 452–453 (1984)
621. Hill, I.D.: Discussion of "A new probability model for determining exact P-values for 2×2 contingency tables when comparing binomial proportions" by W.R. Rice. Biometrics **44**, 14–16 (1988)
622. Hill, I.D., Pike, M.C.: Algorithm 4: TWOBYTWO. Comput. Bull. **9**, 56–63 (1965) [Reprinted in Comput. J. **22**, 87–88 (1979); addenda in Comput. J. **9**, 212 (1966) and **9**, 416 (1967)]

623. Hilton, J.F., Mehta, C.R., Patel, N.R.: An algorithm for conducting exact Smirnov tests. Comput. Stat. Data Anal. **17**, 351–361 (1994)
624. Hinkelmann, K.: Statisticians in History: Oscar Kempthorne: 1919–2000. http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=8. Accessed 12 Apr 2012
625. Hinkelmann, K.: Design and Analysis of Experiments, vol. I: Introduction to Experimental Design. Wiley, New York (1994)
626. Hinkelmann, K.: Remembering Oscar Kempthorne (1919–2000). Stat. Sci. **16**, 169–183 (2001)
627. Hinkley, D.V.: Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. J. Am. Stat. Assoc. **75**, 582–584 (1980)
628. Hinkley, D.V.: R.A. Fisher: Some introductory remarks. In: Feinberg, S.E., Hinkley, D.V. (eds.) R. A. Fisher: An Appreciation, pp. 1–5. Springer, Heidelberg (1980)
629. Hirji, K.F.: Exact Analysis of Discrete Data. Chapman & Hall/CRC, Boca Raton (2006)
630. Hirji, K.F., Johnson, T.D.: A comparison of algorithms for exact analysis of unordered $2 \times K$ contingency tables. Comput. Stat. Data Anal. **21**, 419–429 (1996)
631. Hirji, K.F., Mehta, C.R., Patel, N.R.: Computing distributions for exact logistic regression. J. Am. Stat. Assoc. **82**, 1110–1117 (1987)
632. Hirji, K.F., Tan, S., Elashoff, R.M.: A quasi-exact test for comparing two binomial proportions. Stat. Med. **10**, 1137–1153 (1991)
633. Hitchcock, D.B.: Yates and contingency tables: 75 years later. Elec. J. Hist. Prob. Stat. **5**, 1–14 (2009)
634. Ho, S.T., Chen, L.H.: An L_p bound for the remainder in a combinatorial central limit theorem. Ann. Math. Stat. **2**, 231–249 (1978)
635. Hodges, J.L., Lehmann, E.L.: The efficiency of some non-parametric competitors of the *t*-test. Ann. Math. Stat. **27**, 324–335 (1956)
636. Hoeffding, W.: The large-sample power of tests based on permutations of observations. Ann. Math. Stat. **23**, 169–192 (1952)
637. Hoeffding (Höffding), W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat. **19**, 293–325 (1948)
638. Holford, T.R.: Editorial: Exact methods for categorical data. Stat. Methods Med. Res. **12**, 1 (2003)
639. Hollander, M.: A conversation with Ralph A. Bradley. Stat. Sci. **16**, 75–100 (2000)
640. Hollerith, H.: An electric tabulating system. Quarterly **10**, 238–255 (1889) [Published by the Columbia University School of Mines]
641. Hollerith, H.: The electric tabulating machine. J. R. Stat. Soc. **57**, 678–682 (1894)
642. Holley, J.W., Guilford, J.P.: A note on the g index of agreement. Educ. Psychol. Meas. **4**, 749–753 (1964)
643. Holmes, C.B.: Sample size in psychological research. Percept. Motor Skill. **49**, 283–288 (1979)
644. Holmes, C.B.: Sample size in four areas of psychological research. T. Kan. Acad. Sci. **86**, 76–80 (1983)
645. Holmes, C.B., Holmes, J.R., Famming, J.J.: Sample size in non-APA journals. J. Psychol. **108**, 263–266 (1981)
646. Holschuh, N.: Randomization and design: I. In: Feinberg, S.E., Hinkley, D.V. (eds.) R. A. Fisher: An Appreciation, pp. 35–45. Springer, Heidelberg (1980)
647. Hooper, P.M.: Experimental randomization and the validity of normal-theory inference. J. Am. Stat. Assoc. **84**, 576–586 (1989)
648. Hooton, J.W.L.: Randomization tests: Statistics for experimenters. Comput. Methods Prog. Biomed. **35**, 43–51 (1991)
649. Hope, A.C.A.: A simplified Monte Carlo significance test procedure. J. R. Stat. Soc. B Met. **30**, 582–598 (1968)
650. Horn, S.D.: Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. Biometrics **33**, 237–247 (1977)

651. Hotelling, H.: Review of Statistical Methods for Research Workers. J. Am. Stat. Assoc. **22**, 411–412 (1927)
652. Hotelling, H.: The generalization of Student's ratio. Ann. Math. Stat. **2**, 360–378 (1931)
653. Hotelling, H., Pabst, M.R.: Rank correlation and tests of significance involving no assumption of normality. Ann. Math. Stat. **7**, 29–43 (1936)
654. Householder, A.S., Forsythe, G.E., Germond, H.H. (eds.): Monte Carlo Methods. No. 12 in National Bureau of Standards Applied Mathematics Series. United States Government Printing Office, Washington, DC (1951)
655. Howard, J.V.: The 2 × 2 table: A discussion from a Bayesian viewpoint. Stat. Sci. **13**, 351–367 (1998)
656. Howell, D.C.: Statistical Methods for Psychology, 8th edn. Wadsworth, Belmont (2013)
657. Howell, D.C., Gordon, L.R.: Computing the exact probability of an r by c contingency table with fixed marginal totals. Behav. Res. Methods Instrum. **8**, 317 (1976)
658. Howell, J.R.: Algorithm 87: Permutation generator. Commun. ACM **5**, 209 (1962)
659. Howell, J.R.: Generation of permutations by addition. Math. Comput. **16**, 243–244 (1962)
660. Hsu, L.M., Field, R.: Interrater agreement measures: Comments on kappa$_n$, Cohen's kappa, Scott's π, and Aikin's α. Understand. Stat. **2**, 205–219 (2003)
661. Hu, T.C., Tien, B.N.: Generating permutations with nondistinct items. Am. Math. Mon. **83**, 629–631 (1976)
662. Huang, A., Jin, R., Robinson, J.: Robust permutation tests for two samples. J. Stat. Plan. Infer. **139**, 2631–2642 (2009)
663. Hubbard, R.: Alphabet soup: Blurring the distinctions between p's and α's in psychological research. Theor. Psychol. **14**, 295–327 (2004)
664. Hubbard, R., Bayarri, M.J.: Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing (with discussion). Am. Stat. **57**, 171–182 (2003)
665. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
666. Hubert, L.: Assignment Methods in Combinatorial Data Analysis. Marcel Dekker, New York (1987)
667. Hubert, L.J.: Kappa revisited. Psychol. Bull. **84**, 289–297 (1977)
668. Hubert, L.J.: A general formula for the variance of Cohen's weighted kappa. Psychol. Bull. **85**, 183–184 (1978)
669. Huh, M.H., Jhun, M.: Random permutation testing in multiple linear regression. Commun. Stat. Theor. M. **30**, 2023–2032 (2001)
670. Hurvich, C.M., Tsai, C.L.: Model selection for least absolute deviations regression in small samples. Stat. Prob. Lett. **9**, 259–265 (1990)
671. Hutchinson, T.P.: Kappa muddles together two sources of disagreement: Tetrachoric correlation is preferable. Res. Nurs. Health **16**, 313–315 (1993)
672. Hyman, A.: Charles Babbage: Pioneer of the Computer. Oxford University Press, Oxford (1982)
673. In Memoriam: Dr. Colin White. Yale Univ. News. http://news.yale.edu/2011/03/14/memoriam-dr-colin-white (2011). Accessed 31 Mar 2012
674. Irwin, J.O.: Tests of significance for differences between percentages based on small numbers. Metron **12**, 83–94 (1935)
675. Ives, F.W.: Permutation enumeration: Four new permutation algorithms. Commun. ACM **19**, 68–72 (1976)
676. Iyer, H.K., Berry, K.J., Mielke, P.W.: Computation of finite population parameters and approximate probability values for multi-response randomized block permutations (MRBP). Commun. Stat. Simul. C **12**, 479–499 (1983)
677. Jacobson, J.E.: The Wilcoxon two-sample statistic: Tables and bibliography. J. Am. Stat. Assoc. **58**, 1086–1103 (1963)
678. Jagger, G.: Discussion of "Tests of significance in 2 × 2 tables" by F. Yates. J. R. Stat. Soc. A Gen. **147**, 455 (1984)
679. Janson, S., Vegelius, J.: On generalizations of the g index and the phi coefficient to nominal scales. Multivar. Behav. Res. **14**, 255–269 (1979)

680. Janssen, A., Pauls, T.: How do bootstrap and permutation tests work? Ann. Stat. **31**, 768–806 (2003)
681. Jarrett, T.: On algebraic notation. Trans. Camb. Philos. Soc. **3**, 65–103 (1830)
682. Jastrow, J.: Joseph Jastrow. In: Murchison, C. (ed.) A History of Psychology in Autobiography, vol. I, pp. 135–162. Russell and Russell, New York (1961)
683. Jenkins, N.: W. H. Auden–'Family Ghosts'. Department of English, Stanford University. http://www.stanford.edu/group/auden/cgi-bin/auden (2008). Accessed 6 Feb 2012
684. Jensen, A.R.: Charles E. Spearman: The discoverer of *g*. In: Kimble, G.A., Wertheimer, M. (eds.) Portraits of Pioneers in Psychology, vol. IV, pp. 93–111. Lawrence Erlbaum, Mahwah, NJ (2000)
685. Jeyaratnam, S.: Confidence intervals for the correlation coefficient. Stat. Prob. Lett. **15**, 389–393 (1992)
686. Jiang, W., Kalbfleisch, J.D.: Permutation methods in relative risk regression models. J. Stat. Plan. Infer. **138**, 416–431 (2008)
687. Jin, R., Robinson, J.: Robust permutation tests for one sample. J. Stat. Plan. Infer. **116**, 475–487 (2003)
688. Joe, H.: Extreme probabilities for contingency tables under row and column independence with application to Fisher's exact test. Commun. Stat. Theor. Methods **17**, 3677–3685 (1988)
689. Johnson, E.M.: The Fisher–Yates exact test and unequal sample sizes. Psychometrika **37**, 103–106 (1972)
690. Johnson, N.L.: Theoretical considerations regarding H.R.B. Hack's system of randomization for cross-classifications. Biometrika **45**, 265–266 (1958)
691. Johnson, N.L., Kotz, S. (eds.): Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present. Wiley Series in Probability and Statistics. Wiley, New York (1997)
692. Johnson, N.L., Kotz, S.: Wilks, Samuel Stanley. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 211–212. Wiley, New York (1997)
693. Johnson, S.M.: Generation of permutations by adjacent transposition. Math. Comput. **17**, 282–285 (1963)
694. Johnson, W.D., Mercante, D.E.: Applications of the IML Procedure for Multiple Response Permutation Tests. SAS Users Group International 18, New York. http://www.sascommunity.org/sugi/SUGI93/Sugi-93-173%20Johnson%20Mercante.pdf (1993). Accessed 13 July 2012
695. Johnston, J.E.: Amenity, community, and ranching: Rancher's beliefs, behaviors, and attitudes regarding ranching in the West. Unpublished dissertation, Colorado State University (2006)
696. Johnston, J.E., Berry, K.J., Mielke, P.W.: Permutation tests: Precision in estimating probability values. Percept. Motor Skill. **105**, 915–920 (2007)
697. Johnston, J.E., Berry, K.J., Mielke, P.W.: Quantitative historical methods: A permutation alternative. Hist. Methods **42**, 35–39 (2009)
698. Jolayemi, E.T.: On the measure of agreement between two raters. Biometrical J. **32**, 87–93 (1990)
699. Jonckheere, A.R.: A distribution-free *k*-sample test against ordered alternatives. Biometrika **41**, 133–145 (1954)
700. Jones, L.V. (ed.): The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis, 1949–1964, vol. III. Wadsworth Statistics and Probability Series. Wadsworth & Brooks/Cole, Belmont (1985)
701. Jones, L.V. (ed.): The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis, 1965–1984, vol. IV. Wadsworth Statistics and Probability Series. Wadsworth & Brooks/Cole, Belmont (1986)
702. Joyce, J.: Interview with Bill Joy. Unix Review. http://web.cecs.pdx.edu/~kirkenda/joy84.html (August 1984). Accessed 6 Oct 2012

703. Jung, B.C., Jhun, M., Song, S.H.: A new random permutation test in ANOVA models. Stat. Pap. **48**, 47–62 (2007)
704. Kafadar, K.: In memoriam: John Wilder Tukey June 16, 1915 – July 26, 2000. Technometrics **43**, 251–255 (2001)
705. Kaiser, J.: An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples. Stata J. **7**, 402–412 (2007)
706. Kalish, L.A.: Permutation tests following restricted randomization procedures. Control. Clin. Trials **11**, 147–149 (1990)
707. Kamat, A.R.: A two-sample distribution-free test. Biometrika **43**, 388–387 (1956)
708. Kanellos, M.: Moore says nanoelectronics face tough challenges. http://news.cnet.com/2100-1006_3-5607422.html (9 March 2005). Accessed 16 Jan 2012
709. Kannemann, K.: The exact evaluation of 2-way cross-classifications: An algorithmic solution. Biometrical J. **24**, 157–169 (1982)
710. Kannemann, K.: The exact evaluation of 2-way cross-classifications. Sequel: A fugal algorithm. Biometrical J. **24**, 679–684 (1982)
711. Kaufman, E.H., Taylor, G.D., Mielke, P.W., Berry, K.J.: An algorithm and FORTRAN program for multivariate LAD (ℓ_1 of ℓ_2) regression. Computing **68**, 275–287 (2002)
712. Kean, S.: The Disappearing Spoon. Little, Brown, New York (2010)
713. Kell, H.J., Lubinski, D., Benbow, C.P.: Who rises to the top? Early indicators. Psychol. Sci. **24**, 648–659 (2013)
714. Keller-McNulty, S., Higgins, J.J.: Effect of tail weight and outliers and power and type-I error of robust permutation tests for location. Commun. Stat. Simul. C **16**, 17–35 (1987)
715. Kelley, K.: Petitions requested for new section. Amstat News **419**, 9 (May 2012)
716. Kelly, F.P., Vonder Haar, T.H., Mielke, P.W.: Imagery randomized block analysis (IRBA) applied to the verification of cloud edge detectors. J. Atmos. Ocean. Tech. **6**, 671–679 (1989)
717. Kemp, A.W., Kemp, C.D.: Weldon's dice data revisited. Am. Stat. **45**, 216–222 (1991)
718. Kempthorne, O.: The Design and Analysis of Experiments. Wiley, New York (1952)
719. Kempthorne, O.: The randomization theory of experimental inference. J. Am. Stat. Assoc. **50**, 946–967 (1955)
720. Kempthorne, O.: Some aspects of experimental inference. J. Am. Stat. Assoc. **61**, 11–34 (1966)
721. Kempthorne, O.: Why randomize? J. Stat. Plan. Infer. **1**, 1–25 (1977)
722. Kempthorne, O.: In dispraise of the exact test: Reactions. J. Stat. Plan. Infer. **3**, 199–213 (1979)
723. Kempthorne, O.: Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. J. Am. Stat. Assoc. **75**, 584–587 (1980)
724. Kempthorne, O.: Revisiting the past and anticipating the future. In: David, H.A., David, H.T. (eds.) Statistics: An Appraisal, pp. 31–52. Iowa State University Press, Ames (1984)
725. Kempthorne, O., Doerfler, T.E.: The behaviour of some significance tests under experimental randomization. Biometrika **56**, 231–248 (1969)
726. Kempthorne, O., Zyskind, G., Addelman, S., Throckmorton, T.N., White, R.F.: Constrained randomization. In: Analysis of Variance Procedures, vol. 129, Aeronautical Research Laboratory Reports, chap. VIII, pp. 190–202. Office of Aerospace Research, United States Air Force, Wright-Patterson Air Force Base, Ohio (1961)
727. Kendall, D.G., Bartlett, M.S., Page, T.L.: Jerzy Neyman: 1894–1981. Bull. Lond. Math. Soc. **16**, 160–168 (1984)
728. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**, 81–93 (1938)
729. Kendall, M.G.: The Advanced Theory of Statistics, vol. I. Griffin, London (1943)
730. Kendall, M.G.: The treatment of ties in ranking problems. Biometrika **33**, 239–251 (1945)
731. Kendall, M.G.: The Advanced Theory of Statistics, vol. II. Griffin, London (1946)
732. Kendall, M.G.: Discussion of "The statistical study of infectious diseases" by M. Greenwood. J. R. Stat. Soc. **109**, 103–105 (1946)

733. Kendall, M.G.: The variance of τ when both rankings contain ties. Biometrika **34**, 297–298 (1947)
734. Kendall, M.G.: Rank Correlation Methods. Griffin, London (1948)
735. Kendall, M.G.: Studies in the history of probability and statistics: XI. Daniel Bernoulli on maximum likelihood. Biometrika **48**, 1–18 (1961)
736. Kendall, M.G.: Rank Correlation Methods, 3rd edn. Griffin, London (1962)
737. Kendall, M.G.: Statistical inference in the light of the theory of the electronic computer. Rev. Int. Stat. Inst. **34**, 1–12 (1966)
738. Kendall, M.G.: The history and future of statistics. In: Bancroft, T.A. (ed.) Statistical Papers in Honor of George W. Snedecor, pp. 193–210. Iowa State University Press, Ames (1972)
739. Kendall, M.G., Babington Smith, B.: The problem of m rankings. Ann. Math. Stat. **10**, 275–287 (1939)
740. Kendall, M.G., Babington Smith, B.: Tables of random sampling numbers. In: Tracts for Computers, vol. 24. Cambridge University Press, Cambridge (1939)
741. Kendall, M.G., Babington Smith, B.: On the method of paired comparisons. Biometrika **31**, 324–345 (1940)
742. Kendall, M.G., Buckland, W.R.: A Dictionary of Statistical Terms. Oliver and Boyd, Edinburgh (1957)
743. Kendall, M.G., Doig, A.G.: Bibliography of Statistical Literature: 1950–1958. Oliver and Boyd, Edinburgh (1962)
744. Kendall, M.G., Doig, A.G.: Bibliography of Statistical Literature: 1940–1949. Hafner, New York (1965)
745. Kendall, M.G., Doig, A.G.: Bibliography of Statistical Literature: Pre 1940. Oliver and Boyd, Edinburgh (1968)
746. Kendall, M.G., Kendall, S.F.H., Babington Smith, B.: The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. Biometrika **30**, 251–273 (1939)
747. Kendall, M.G., Plackett, R.L. (eds.): Studies in the History of Statistics and Probability, vol. II. Griffin, London (1977)
748. Kennedy, P.E.: Randomization tests in econometrics. J. Bus. Econ. Stat. **13**, 85–94 (1995)
749. Kennedy, P.E., Cade, B.S.: Randomization tests for multiple regression. Commun. Stat. Simul. C **25**, 923–936 (1996)
750. Ker, M.: Issues in the use of kappa. Invest. Radiol. **26**, 78–83 (1991)
751. Kiang, L.Y.: Charles Stein: The invariant, the direct and the "pretentious". In: Kiang, L.Y. (ed.) Creative Minds, Charmed Lives: Interviews at Institute for Mathematical Sciences, National University of Singapore, pp. 282–287. World Scientific, Singapore (2010)
752. Kiang, L.Y. (ed.): Creative Minds, Charmed Lives: Interviews at Institute for Mathematical Sciences, National University of Singapore. World Scientific, Singapore (2010)
753. Kiernan, D.: The Girls of Atomic City: The Untold Story of the Women Who Helped Win World War II. Touchstone, New York (2013)
754. Killion, R.A., Zahn, D.A.: Bibliography of contingency table literature: 1900 to 1974. Int. Stat. Rev. **44**, 71–112 (1976)
755. Kim, A.: Wilhelm Maximilian Wundt. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Fall, 2008 edn. Stanford University. http://plato.stanford.edu/arhives/fall2008/entries/wilhelm-wundt/ (2008). Accessed 10 Oct 2011
756. Kim, M.J., Nelson, C.R., Startz, R.: Mean revision in stock prices? A reappraisal of the empirical evidence. Rev. Econ. Stud. **58**, 515–528 (1991)
757. Kingman, A.: Beyond weighted kappa when evaluating examiner agreement for ordinal responses. J. Dent. Res. **81**, A219 (2002)
758. Kingman, J.F.C.: Uses of exchangeability. Ann. Prob. **6**, 183–197 (1978) [Abraham Wald memorial lecture delivered in August 1977 in Seattle, Washington]
759. Klotz, J.: Nonparametric tests for scale. Ann. Math. Stat. **33**, 498–512 (1962)
760. Klotz, J., Teng, J.: One-way layout for counts and the exact enumeration of the Kruskal–Wallis H distribution with ties. J. Am. Stat. Assoc. **72**, 165–169 (1977)

761. Knijnenburg, T.A., Wessels, L.F.A., Reinders, M.J.T., Shmulevich, I.: Fewer permutations, more accurate *P*-values. Intell. Syst. Mol. Biol. **25**, i161–i168 (2009)
762. Knuth, D.E.: The Art of Computer Programming: Seminumerical Algorithms, vol. II. Addison-Wesley, Reading (1969)
763. Knuth, D.E.: The Art of Computer Programming: Seminumerical Algorithms, vol. II, 2nd edn. Addison-Wesley, Reading (1981)
764. Ko, C.W., Ruskey, F.: Generating permutations of a bag by interchanges. Inform. Process. Lett. **41**, 263–269 (1992)
765. Koenker, R., Bassett, G.: Tests of linear hypotheses and l_1 estimation. Econometrica **50**, 1577–1583 (1982)
766. Kolmogorov, A.N.: Sulla determinazione empirica di una legge di distribuzione (On the empirical distribution of a distribution). Inst. Ital. Attuari, Giorn. **4**, 83–91 (1933)
767. Kotz, S., Johnson, N.L. (eds.): Breakthroughs in Statistics: Foundations and Basic Theory, vol. I. Springer Series in Statistics. Springer, New York (1992)
768. Kraft, C.A., van Eeden, C.: A Nonparametric Introduction to Statistics. Macmillan, New York (1968)
769. Kramer, M.S., Feinstein, A.R.: Clinical biostatistics: LIV. The biostatistics of concordance. Clin. Pharm. Therap. **29**, 111–123 (1981)
770. Kramp, C.: Éléments d'arithmétique universelle. Hansen, Cologne (1808)
771. Kreiner, S.: Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. Scand. J. Stat. **14**, 97–112 (1987)
772. Krippendorff, K.: Bivariate agreement coefficients for reliability of data. In: Borgatta, E.G. (ed.) Sociological Methodology, pp. 139–150. Jossey-Bass, San Francisco (1970)
773. Kromrey, J.D., Chason, W.M., Blair, R.C.: PERMUTE: A SAS algorithm for permutation testing. Appl. Psychol. Meas. **16**, 64 (1992)
774. Kroonenberg, P.M., Verbeek, A.: Comment on "The exact evaluation of 2-way cross-classifications" by K. Kannemann. Biometrical J. **27**, 719–720 (1985)
775. Krüger, L., Daston, L., Heidelberger, M. (eds.): The Probabilistic Revolution, vol. I: Ideas in History. MIT Press, Cambridge (1987)
776. Kruskal, W.H.: Historical notes on the Wilcoxon unpaired two-sample test. J. Am. Stat. Assoc. **52**, 356–360 (1957)
777. Kruskal, W.H.: Discussion of the papers of Messrs. Anscombe and Daniel. Technometrics **2**, 157–158 (1960)
778. Kruskal, W.H.: The significance of Fisher: A review of *R.A. Fisher: The Life of a Scientist* by J.F. Box. J. Am. Stat. Assoc. **75**, 1019–1030 (1980)
779. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. **47**, 583–621 (1952) [Erratum: J. Am. Stat. Assoc. **48**, 907–911 (1953)]
780. Kulinskaya, E.: Large sample results for permutation tests of association. Commun. Stat. Theor. Methods **23**, 2939–2963 (1994)
781. Kundel, H.L., Polansky, M.: Measurement of observer agreement. Radiology **228**, 303–308 (2003)
782. Kurtz, T.E.: Basic. ACM SIGPLAN Notices **13**, 103–118 (1978)
783. Kurtzberg, J.: Algorithm 94: Combination. Commun. ACM **5**, 5 (1962)
784. Kwok, R.: Salvage job. Sci. News **183**, 20–24 (23 February 2013)
785. Lachin, J.M.: Properties of randomization in clinical trials: Foreword. Control. Clin. Trials **9**, 287–288 (1988)
786. Lachin, J.M.: Properties of simple randomization in clinical trials. Control. Clin. Trials **9**, 312–326 (1988)
787. Lachin, J.M.: Statistical properties of randomization in clinical trials. Control. Clin. Trials **9**, 289–311 (1988)
788. Lachin, J.M., Matts, J.P., Wei, L.J.: Randomization in clinical trials: Conclusions and recommendations. Control. Clin. Trials **9**, 312–326 (1988)
789. LaFleur, B.J., Greevy, R.A.: Introduction to permutation and resampling-based hypothesis tests. J. Clin. Child Adolesc. **38**, 286–294 (2009)

790. Lahiri, S.N.: Resampling Methods for Dependent Data. Springer, New York (2003)
791. Lambert, D.: Robust two-sample permutation tests. Ann. Stat. **13**, 606–625 (1985)
792. Lancaster, H.O.: The combination of probabilities arising from data in discrete distributions. Biometrika **36**, 370–382 (1949)
793. Lancaster, H.O.: The sex ratios in sibships with special reference to Geissler's data. Ann. Hum. Genet. **15**, 153–158 (1949)
794. Lancaster, H.O.: Significance test in discrete distributions. J. Am. Stat. Assoc. **56**, 223–234 (1961) [Corrigendum: J. Am. Stat. Assoc. **57**, 919 (1962)]
795. Lance, C.E.: More statistical and methodological myths and urban legends. Organ. Res. Methods **14**, 279–286 (2011)
796. Landis, J.R., Koch, G.G.: The measurement of observer agreement for ordinal data. Biometrics **33**, 159–174 (1977)
797. Lane, D.A.: Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. J. Am. Stat. Assoc. **75**, 589–590 (1980)
798. Langbehn, D.R.: Comment on "Why permutation tests are superior to t and F tests in biomedical research" by J. Ludbrook and H.A.F. Dudley. Am. Stat. **54**, 85 (2000)
799. Langdon, Jr., G.G.: An algorithm for generating permutations. Commun. ACM **10**, 5 (1967)
800. Langdon, Jr., G.G.: Generating permutations by nested cycling. Commun. ACM **11**, 392 (1968)
801. Lange, J.: Crime as Destiny: A Study of Criminal Twins. Allen & Unwin, London (1931) [Translated by C. Haldane]
802. Laplace, P.: Mécanique Céleste, vol. II. Hilliard, Gray, Little, and Wilkins, Boston (1832) [English translation, with notes and commentary, by Nathaniel Bowditch]
803. Laster, L.L.: Permutation tests: Fisher's (1925) test of a wider hypothesis. J. Dent. Res. **77**, 906 (1998) [Abstract of a presentation at the Symposium on Behavioral Sciences and Health Services Research, International Association of Dental Research, June 1998 in Nice, France]
804. Latscha, R.: Tests of significance in a 2 × 2 contingency table: Extension of Finney's table. Biometrika **40**, 74–86 (1953)
805. Lazarsfeld, P.F.: Notes on the history of quantification in sociology — Trends, sources, and problems. Isis **52**, 277–333 (1961)
806. Leach, C.: Introduction to Statistics: A Nonparametric Approach for the Social Sciences. Wiley, New York (1979)
807. L'Ecuyer, P.: Uniform random number generation. Ann. Oper. Res. **53**, 77–120 (1994)
808. Ledermann, W.: Walter Ledermann: Encounters of a Mathematician. http://www-history.mcs.st.andrews.ac.uk/Ledermann/Ch7.html (2000). Accessed 6 Feb 2012
809. Lee, T.J., Pielke, R.A., Mielke, P.W.: Modeling the clear-sky surface energy budget during FIFE 1987. J. Geophys. Res. **100**, 25,585–25,593 (1995)
810. Legendre, P.: Species associations: The Kendall coefficient of concordance revisited. J. Agric. Biol. Environ. Sci. **10**, 226–245 (2005)
811. Legendre, P., Gallagher, E.D.: Ecologically meaningful transformations for ordination of species data. Oecologia **129**, 271–280 (2001)
812. Lehmann, E.L.: Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco (1975)
813. Lehmann, E.L.: "Student" and small-sample theory. Stat. Sci. **14**, 418–426 (1999)
814. Lehmann, E.L.: Reminiscences of a Statistician: The Company I Kept. Springer, New York (2008)
815. Lehmann, E.L.: Parametrics vs. nonparametrics: Two alternative methodologies. J. Nonparametr. Stat. **21**, 397–405 (2009)
816. Lehmann, E.L.: Fisher, Neyman, and the Creation of Classical Statistics. Springer, New York (2011)
817. Lehmann, E.L., Reid, C.: Jerzy Neyman, 1894–1981. Am. Stat. 161–162 (1982)
818. Lehmann, E.L., Stein, C.M.: On the theory of some non-parametric hypotheses. Ann. Math. Stat. **20**, 28–45 (1949)

819. Gottfried Wilhelm Leibniz invents the binary system. The Centre for Computing History. http://www.computinghistory.org.uk/det/5913/Gottfried%20Wilhelm%20Leibniz%20invents%20the%20Binary%20System (2012). Accessed 12 Sept 2012
820. Lenstra, J.K.: Recursive algorithms for enumerating subsets, lattice-points, combinations and permutations. CWI Tech. Rep. BW 28/73, Stichting Mathematisch Centrum, Amsterdam (1973)
821. Leslie, P.H.: A simple method of calculating the exact probability in 2×2 contingency tables with small marginal totals. Biometrika **42**, 522–523 (1955)
822. Levin, B., Robbins, H.: Urn models for regression analysis, with applications to employment discrimination studies. Law Contemp. Probl. **46**, 247–267 (1983)
823. Levin, R.C.: A Yale pioneer: The freshman address. Yale Alum. Mag., 1–4. http://www.yalealumnimagazine.com/issues/2009_11/levin5962.html (November/December 2009). Accessed 14 Mar 2012
824. Levy, D.: George B. Dantzig, operations research professor, dies at 90. http://news,stanford.edu/news/2005/may25/dantzigobit-052505.html (May 2005). Accessed 24 December 2011
825. Light, R.J.: Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol. Bull. **76**, 365–377 (1971)
826. Linders, F.J., Hurlin, R.G.: Personal notes. J. Am. Stat. Assoc. **30**, 751–754 (1935)
827. Lindgren, F., Hansen, B., Karcher, W., Sjostrom, M., Eriksson, L.: Model validation by permutation tests: Applications to variable selection. J. Chemometr. **10**, 521–532 (1995)
828. Lindley, D.V.: Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. J. Am. Stat. Assoc. **75**, 589–590 (1980)
829. Lindley, D.V.: A Bayesian lady tasting tea. In: David, H.A., David, H.T. (eds.) Statistics: An Appraisal, pp. 455–479. Iowa State University Press, Ames (1984)
830. Lindley, D.V.: Professor George A. Barnard, 1915–2002. Statistician **52**, 231–234 (2003)
831. Lindley, D.V.: George Barnard: Anti-establishment statistician concerned with quality control. The Guardian. http://www.guardian.co.uk/news/2002/aug/09/guardianobituaries.highereducation (9 August 2002). Accessed 2 Oct 2012
832. Lipski, Jr., W.: More on permutation generation methods. Computing **23**, 357–365 (1979)
833. Litchfield, Jr., J.T., Wilcoxon, F.: The rank correlation method. Anal. Chem. **27**, 299–300 (1955)
834. Littell, R.C., Folks, J.L.: Asymptotic optimality of Fisher's method of combining independent tests. J. Am. Stat. Assoc. **66**, 802–806 (1971)
835. Littell, R.C., Folks, J.L.: Asymptotic optimality of Fisher's method of combining independent tests: II. J. Am. Stat. Assoc. **68**, 193–194 (1973)
836. Little, R.J.A.: Testing the equality of two independent binomial proportions. Am. Stat. **43**, 283–288 (1989)
837. Liu, C.N., Tang, D.T.: Algorithm 452: Enumerating combinations of m out of n objects. Commun. ACM **16**, 485 (1973)
838. Lloyd, C.J.: Doubling the one-sided P-value in testing independence in 2×2 tables against a two-sided alternative. Stat. Med. **7**, 1297–1306 (1988)
839. Lomb, N.: Transit of Venus: 1631 to the Present. Powerhouse Museum, Sydney (2011)
840. Long, M.A., Berry, K.J., Mielke, P.W.: A note on tests of significance for multiple regression coefficients. Psychol. Rep. **100**, 339–345 (2007)
841. Long, M.A., Berry, K.J., Mielke, P.W.: Tetrachoric correlation: A permutation alternative. Educ. Psychol. Meas. **69**, 429–437 (2009)
842. Look, B.C.: Gottfried Wilhelm Leibniz. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Fall 2012 edn. Stanford University. http://plato.stanford.edu/archives/fall2012/entries/leibniz/ (2012). Accessed 12 Sept 2012
843. Lotto, G.: On the generation of all possible stepwise combinations. Math. Comput. **16**, 214–243 (1962)
844. Lovie, A.D.: Who discovered Spearman's rank correlation? Br. J. Math. Stat. Psychol. **48**, 255–269 (1995)

845. Lovie, P.: Charles Edward Spearmen F.R.S. 1863–1945. A commemoration of the 50th anniversary of his death. Br. J. Math. Stat. Psychol. **48**, 209–210 (1995)

846. Lovie, P., Lovie, A.D.: Charles Edward Spearman, F.R.S. (1863–1945). Notes Rec. R. Soc. Lond. **50**, 75–88 (1996)

847. Lovie, S., Lovie, P.: Commentary: Charles Spearman and correlation: A commentary on 'The proof and measurement of association between two things'. Int. J. Epidemiol. **39**, 1151–1153 (2010)

848. Lowther, A.W.G.: Report on the Excavation of the Roman Structure at Rothamsted Experimental Station, Harpenden. St. Albans Hert. Archit. Arch. Soc. (1937)

849. Ludbrook, J.: Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. Clin. Exp. Pharmacol. Physiol. **21**, 673–686 (1994)

850. Ludbrook, J.: Issues in biomedical statistics: Comparing means by computer-intensive tests. Aust. N.Z. J. Surg. **65**, 812–819 (1995)

851. Ludbrook, J.: The Wilcoxon–Mann–Whitney test condemned. Br. J. Surg. **83**, 136–137 (1996)

852. Ludbrook, J.: Statistical techniques for comparing measures and methods of measurement: A critical review. Clin. Exp. Pharmacol. Physiol. **29**, 527–536 (2002)

853. Ludbrook, J.: Outlying observations and missing values: How should they be handled? Clin. Exp. Pharmacol. Physiol. **35**, 670–678 (2008)

854. Ludbrook, J., Dudley, H.A.F.: Issues in biomedical statistics: Analyzing 2 × 2 tables of frequencies. Aust. N. Z. J. Surg. **64**, 780–787 (1994)

855. Ludbrook, J., Dudley, H.A.F.: Issues in biomedical statistics: Statistical inference. Aust. N. Z. J. Surg. **64**, 630–636 (1994)

856. Ludbrook, J., Dudley, H.A.F.: Why permutation tests are superior to t and F tests in biomedical research. Am. Stat. **52**, 127–132 (1998)

857. Ludbrook, J., Dudley, H.A.F.: Discussion of "Why permutation tests are superior to t and F tests in biomedical research" by J. Ludbrook and H.A.F. Dudley. Am. Stat. **54**, 87 (2000)

858. Lunneborg, C.E.: Data Analysis by Resampling: Concepts and Applications. Duxbury, Pacific Grove (2000)

859. Lush, J.L.: Early statistics at Iowa State University. In: Bancroft, T.A. (ed.) Statistical Papers in Honor of George W. Snedecor, pp. 211–226. Iowa State University Press, Ames (1972)

860. Lydekker, J.W.: The Cressy family of Rothamsted. St. Albans Hert. Archit. Arch. Soc. (1937)

861. Lyons, D.: In race for fastest computer, China outpaces U.S. Newsweek **158**, 57–59 (5 December 2011)

862. Macdonell, W.R.: On criminal anthropometry and the identification of criminals. Biometrika **1**, 177–227 (1902)

863. MacKenzie, D.: Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge. Edinburgh University Press, Edinburgh (1981)

864. Maclure, M., Willett, W.C.: Misinterpretation and misuse of the kappa statistic. Am. J. Epidemiol. **126**, 161–169 (1987)

865. MacMahon, P.A.: Combinatory Analysis, vol. II. Cambridge University Press, Cambridge (1916)

866. MacNeill, I.: A conversation with David J. Finney. Stat. Sci. **8**, 187–201 (1993)

867. Mahalanobis, P.C.: Auxiliary tables for Fisher's z-test in analysis of variance. Indian J. Agric. Sci. **2**, 679–693 (1932)

868. Mahalanobis, P.C.: Professor Ronald Aylmer Fisher. Sankhyā **4**, 265–272 (1938)

869. Mahanti, S.: John Burdon Sanderson Haldane: The ideal of a polymath. Vigyan Prasar Science Portal. http://www.vigyanprasar.gov.in/scientists/JBSHaldane.htm (2007). Accessed 20 Jan 2012

870. Maisel, M., Smart, L.: Admiral Grace Murray Hopper: Pioneer computer scientist. http://www.sdsc.edu/ScienceWomen/hopper.html (1997). Accessed 13 Mar 2012

871. Mallows, C.L. (ed.): The Collected Works of John W. Tukey: More Mathematical, 1938–1984, vol. VI. Wadsworth Statistics and Probability Series. Wadsworth & Brooks/Cole, Pacific Grove (1990)

872. Mallows, C.L.: Comment on “Why permutation tests are superior to t and F tests in biomedical research” by J. Ludbrook and H.A.F. Dudley. Am. Stat. **54**, 86–87 (2000)
873. Maltz, M.D.: Deviating from the mean: The declining significance of significance. J. Res. Crime Delinq. **31**, 434–463 (1994)
874. Management: Du Pont educates engineers. Chem. Eng. News **36**, 38–39 (1958)
875. Manly, B.F.J.: Randomization and Monte Carlo Methods in Biology. Chapman & Hall, London (1991)
876. Manly, B.F.J.: Randomization and Monte Carlo Methods in Biology, 2nd edn. Chapman & Hall, London (1997)
877. Manly, B.F.J.: Randomization, Bootstrap and Monte Carlo Methods in Biology, 3rd edn. Chapman & Hall/CRC, Boca Raton (2007)
878. Manly, B.F.J., Francis, R.I.C.: Analysis of variance by randomization when variances are unequal. Aust. N. Z. J. Stat. **41**, 411–429 (1999)
879. Mann, H.B.: Nonparametric tests against trend. Econometrica **13**, 245–259 (1945)
880. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**, 50–60 (1947)
881. Mantel, N.: The detection of disease clustering and a generalized regression approach. Cancer Res. **27**, 209–220 (1967)
882. Mantel, N.: 361: Approaches to a health research occupancy problem. Biometrics **30**, 355–362 (1974)
883. Mantel, N.: Comment on “Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables” by W.J. Conover. J. Am. Stat. Assoc. **69**, 378–380 (1974)
884. Mantel, N.: Discussion of “Tests of significance in 2 × 2 tables” by F. Yates. J. R. Stat. Soc. A Gen. **147**, 457–458 (1984)
885. Mantel, N.: Comment on “Yates’s correction for continuity and the analysis of 2 × 2 contingency tables” by M.G. Haviland. Stat. Med. **9**, 369–370 (1990)
886. Mantel, N., Greenhouse, S.W.: What is the continuity correction? Am. Stat. **22**, 27–30 (1968)
887. Mantel, N., Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. **22**, 719–748 (1959)
888. Mantel, N., Pasternack, B.S.: A class of occupancy problems. Am. Stat. **22**, 23–24 (1968)
889. Mantel, N., Valand, R.S.: A technique of nonparametric multivariate analysis. Biometrics **26**, 547–558 (1970)
890. March, D.L.: Algorithm 434: Exact probabilities for $R \times C$ contingency tables. Commun. ACM **15**, 991–992 (1972)
891. Marcuson, R.: A matrix formulation for nonparametric permutation tests. Biometrical J. **38**, 887–891 (1996)
892. Mardia, K.V.: Obituary: Professor B. L. Welch. J. R. Stat. Soc. A Stat. **153**, 253–254 (1990)
893. Marks, H.M.: The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990. Cambridge University Press, Cambridge (1997)
894. Marriott, F.H.C.: Barnard’s Monte Carlo tests: How many simulations? J. R. Stat. Soc. C Appl. Stat. **28**, 75–77 (1979)
895. Marsaglia, G., Bray, T.A.: One-line random number generators and their use in combinations. Commun. ACM **11**, 737–759 (1968)
896. Marszalek, J.M., Barber, C., Kohlhart, J., Holmes, C.B.: Sample size in psychological research over the past 30 years. Percept. Motor Skill. **112**, 331–348 (2011)
897. Martín Andrés, A.: Comments on “Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations” by I. Campbell. Stat. Med. **27**, 1791–1795 (2008)
898. Martín Andrés, A., Herranz Tejedor, I., Luna del Castillo, J.D.: Optimal correction for continuity in the chi-squared test in 2 × 2 tables. Commun. Stat. Simul. C **21**, 1077–1101 (1992)
899. Martín Andrés, A., Herranz Tejedor, I.H.: Is Fisher’s exact test very conservative? Comput. Stat. Data Anal. **19**, 579–591 (1995)

900. Martín Andrés, A., Luna del Castillo, J.D.: Comment on "Sensitivity of Fisher's exact test to minor perturbations in 2 × 2 contingency tables" by W.D. Dupont. Stat. Med. **8**, 243–245 (1989)
901. Martín Andrés, A., Marzo, P.F.: Delta: A new measure of agreement between two raters. Br. J. Math. Stat. Psychol. **57**, 1–19 (2004)
902. Martín Andrés, A., Marzo, P.F.: Chance-corrected measures of reliability and validity in $k \times k$ tables. Stat. Methods Med. Res. **14**, 473–492 (2005)
903. Martín Andrés, A., Sánchez Quevedo, M.J., Tapia García, J.M., Silva Mato, A.: On the validity condition of the chi-squared test in 2 × 2 tables. Test **14**, 1–30 (2005)
904. Mathew, T., Nordström, K.: Least squares and least absolute deviation procedures in approximately linear models. Stat. Prob. Lett. **16**, 153–158 (1993)
905. Matsumoto, M., Nishimura, T.: Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. S. **8**, 3–30 (1998)
906. Matts, J.P., Lachin, J.M.: Properties of permuted-block randomization in clinical trials. Control. Clin. Trials **9**, 327–344 (1988)
907. Maxwell, A.E.: Coefficients of agreement between observers and their interpretation. Br. J. Psychiatr. **130**, 79–83 (1977)
908. May, R.B., Hunter, M.A.: Some advantages of permutation tests. Can. Psychol. **34**, 401–407 (1993)
909. May, S.M.: Modelling observer agreement: An alternative to kappa. J. Clin. Epidemiol. **47**, 1315–1324 (1994)
910. Dr Donal McCarthy MSc, PhD (1957–1967). http://unstats.un.org/unsd/wsd/docs/Ireland_wsd_Former_DGs_bios.pdf. Accessed 20 Dec 2011
911. McCarthy, M.D.: On the application of the z-test to randomized blocks. Ann. Math. Stat. **10**, 337–359 (1939)
912. McDonald, L.: The Woman Founders of the Social Sciences. Carleton University Press, Ottawa (1994)
913. McDonald, L.L., Davis, B.M., Milliken, G.A.: A nonrandomized unconditional test for comparing two proportions in 2 × 2 contingency tables. Technometrics **19**, 145–157 (1977)
914. McHugh, R.B.: Comment on "Scales and statistics: Parametric and nonparametric" by N.H. Anderson. Psychol. Bull. **60**, 350–355 (1963)
915. McKinney, W.P., Young, M.J., Hartz, A., Bi-Fong Lee, M.: The inexact use of Fisher's exact test in six major medical journals. J. Am. Med. Assoc. **261**, 3430–3433 (1989)
916. McNemar, Q.: Note on the sampling error of the differences between correlated proportions and percentages. Psychometrika **12**, 153–157 (1947)
917. McQueen, G.: Long-horizon mean-reverting stock priced revisited. J. Financ. Quant. Anal. **27**, 1–17 (1992)
918. Mehes, A.: Alvan Feinstein dies at age 75. Yale Daily News. http://www.yaledailynews.com/news/2001/oct/31/alvan-feinstein-dies-at-age-75/ (31 October 2001). Accessed 9 Apr 2012
919. Mehta, C.R., Patel, N.R.: A network algorithm for the exact treatment of the $2 \times k$ contingency table. Commun. Stat. Simul. C **9**, 649–664 (1980)
920. Mehta, C.R., Patel, N.R.: A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Stat. Assoc. **78**, 427–434 (1983)
921. Mehta, C.R., Patel, N.R.: Algorithm 643: FEXACT. A FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. ACM Trans. Math. Software **12**, 154–161 (1986)
922. Mehta, C.R., Patel, N.R.: A hybrid algorithm for Fisher's exact test in unordered $r \times c$ contingency tables. Commun. Stat. Theor. Methods **15**, 387–403 (1986)
923. Mehta, C.R., Patel, N.R., Gray, R.: On computing an exact confidence interval for the common odds ratio in several 2 × 2 contingency tables. J. Am. Stat. Assoc. **80**, 969–973 (1985)
924. Mehta, C.R., Patel, N.R., Senchaudhuri, P.: Exact power and sample-size computations for the Cochran–Armitage trend test. Biometrics **54**, 1615–1621 (1998)

925. Meinert, C.: Clinical Trials: Design, Conduct and Analysis. Oxford University Press, New York (1986)
926. Metropolis, N.: The beginning of the Monte Carlo method. Los Alamos Sci. **15**, 125–130 (1987)
927. Metropolis, N., Ulam, S.: The Monte Carlo method. J. Am. Stat. Assoc. **44**, 335–341 (1949)
928. Mewhort, D.J.K., Johns, B.T., Kelly, M.: Applying the permutation test to factorial designs. Behav. Res. Methods **42**, 366–372 (2010)
929. Meyer, G.J.: Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. Psychol. Assess. **9**, 480–489 (1997)
930. Micceri, T.: The unicorn, the normal curve, and other improbable creatures. Psychol. Bull. **105**, 156–166 (1989)
931. Mielke, P.W.: Note on some squared rank tests with existing ties. Technometrics **9**, 312–314 (1967)
932. Mielke, P.W.: Asymptotic behavior of two-sample tests based on powers of ranks for detecting scale and location alternatives. J. Am. Stat. Assoc. **67**, 850–854 (1972)
933. Mielke, P.W.: Squared rank test appropriate to weather modification cross-over design. Technometrics **16**, 13–16 (1974)
934. Mielke, P.W.: Convenient Beta distribution likelihood techniques for describing and comparing meteorological data. J. Appl. Meterol. **14**, 985–990 (1975)
935. Mielke, P.W.: Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique. Biometrics **34**, 277–282 (1978)
936. Mielke, P.W.: On asymptotic non-normality of null distributions of MRPP statistics. Commun. Stat. Theor. Methods **8**, 1541–1550 (1979) [Errata: Commun. Stat. Theor. Methods **10**, 1795 (1981) and **11**, 847 (1982)]
937. Mielke, P.W.: Goodman–Kruskal tau and gamma. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. III, pp. 446–449. Wiley, New York (1983)
938. Mielke, P.W.: Meteorological applications of permutation techniques based on distance functions. In: Krishnaiah, P.R., Sen, P.K. (eds.) Handbook of Statistics, vol. IV, pp. 813–830. North-Holland, Amsterdam (1984)
939. Mielke, P.W.: Geometric concerns pertaining to applications of statistical tests in the atmospheric sciences. J. Atmos. Sci. **42**, 1209–1212 (1985)
940. Mielke, P.W.: Multi-response permutation procedures. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. V, pp. 724–727. Wiley, New York (1985)
941. Mielke, P.W.: Non-metric statistical analyses: Some metric alternatives. J. Stat. Plan. Infer. **13**, 377–387 (1986)
942. Mielke, P.W.: L_1, L_2 and L_∞ regression models: Is there a difference? J. Stat. Plan. Infer. **16**, 430 (1987)
943. Mielke, P.W.: The application of multivariate permutation methods based on distance functions in the earth sciences. Earth Sci. Rev. **31**, 55–71 (1991)
944. Mielke, P.W., Berry, K.J.: An extended class of matched pairs tests based on powers of ranks. Psychometrika **41**, 89–100 (1976)
945. Mielke, P.W., Berry, K.J.: A extended class of permutation techniques for matched pairs. Commun. Stat. Theor. Methods **11**, 1197–1207 (1982)
946. Mielke, P.W., Berry, K.J.: Asymptotic clarifications, generalizations, and concerns regarding an extended class of matched pairs tests based on powers of ranks. Psychometrika **48**, 483–485 (1983)
947. Mielke, P.W., Berry, K.J.: Non-asymptotic inferences based on the chi-square statistic for r by c contingency tables. J. Stat. Plan. Infer. **12**, 41–45 (1985)
948. Mielke, P.W., Berry, K.J.: Cumulant methods for analyzing independence of r-way contingency tables and goodness-of-fit frequency data. Biometrika **75**, 790–793 (1988)
949. Mielke, P.W., Berry, K.J.: Fisher's exact probability test for cross-classification tables. Educ. Psychol. Meas. **52**, 97–101 (1992)
950. Mielke, P.W., Berry, K.J.: Exact goodness-of-fit probability tests for analyzing categorical data. Educ. Psychol. Meas. **53**, 707–710 (1993)

951. Mielke, P.W., Berry, K.J.: Permutation tests for common locations among samples with unequal variances. J. Educ. Behav. Stat. **19**, 217–236 (1994)
952. Mielke, P.W., Berry, K.J.: Nonasymptotic inferences based on Cochran's Q test. Percept. Motor Skill. **81**, 319–322 (1995)
953. Mielke, P.W., Berry, K.J.: Exact probabilities for first-order, second-order, and third-order interactions in $2 \times 2 \times 2 \times 2$ contingency tables. Educ. Psychol. Meas. **56**, 843–847 (1996)
954. Mielke, P.W., Berry, K.J.: An exact solution to an occupancy problem: A useful alternative to Cochran's Q test. Percept. Motor Skill. **82**, 91–95 (1996)
955. Mielke, P.W., Berry, K.J.: Nonasymptotic probability values for Cochran's Q statistic: A FORTRAN 77 program. Percept. Motor Skill. **82**, 303–306 (1996)
956. Mielke, P.W., Berry, K.J.: Permutation-based multivariate regression analysis: The case for least sum of absolute deviations regression. Ann. Oper. Res. **74**, 259–268 (1997)
957. Mielke, P.W., Berry, K.J.: Permutation covariate analyses of residuals based on Euclidean distance. Psychol. Rep. **81**, 795–802 (1997)
958. Mielke, P.W., Berry, K.J.: Multivariate tests for correlated data in completely randomized designs. J. Educ. Behav. Stat. **24**, 109–131 (1999)
959. Mielke, P.W., Berry, K.J.: Euclidean distance based permutation methods in atmospheric science. Data Min. Knowl. Disc. **4**, 7–27 (2000)
960. Mielke, P.W., Berry, K.J.: The Terpstra–Jonckheere test for ordered alternatives: Randomized probability values. Percept. Motor Skill. **91**, 447–450 (2000)
961. Mielke, P.W., Berry, K.J.: Permutation Methods: A Distance Function Approach. Springer, New York (2001)
962. Mielke, P.W., Berry, K.J.: Categorical independence tests for large sparse R-way contingency tables. Percept. Motor Skill. **95**, 606–610 (2002)
963. Mielke, P.W., Berry, K.J.: Multivariate multiple regression analyses: A permutation method for linear models. Psychol. Rep. **91**, 3–9 (2002)
964. Mielke, P.W., Berry, K.J.: Multivariate multiple regression prediction models: A Euclidean distance approach. Psychol. Rep. **92**, 763–769 (2003)
965. Mielke, P.W., Berry, K.J.: Permutation Methods: A Distance Function Approach, 2nd edn. Springer, New York (2007)
966. Mielke, P.W., Berry, K.J.: Two-sample multivariate similarity permutation comparison. Psychol. Rep. **100**, 257–262 (2007)
967. Mielke, P.W., Berry, K.J.: A note on Cohen's weighted kappa coefficient of agreement with linear weights. Stat. Methodol. **6**, 439–446 (2009)
968. Mielke, P.W., Berry, K.J., Brier, G.W.: Application of multi-response permutation procedures for examining seasonal changes in monthly mean sea-level pressure patterns. Mon. Weather Rev. **109**, 120–126 (1981)
969. Mielke, P.W., Berry, K.J., Brockwell, P.J., Williams, J.S.: A class of nonparametric tests based on multi-response permutation procedures. Biometrika **68**, 720–724 (1981)
970. Mielke, P.W., Berry, K.J., Eighmy, J.L.: A permutation procedure for comparing archaeomagnetic polar directions. In: Eighmy, J.L., Sternberg, R.S. (eds.) Archaeomagnetic Dating, pp. 102–108. University of Arizona Press, Tucson (1991)
971. Mielke, P.W., Berry, K.J., Johnson, E.S.: Multi-response permutation procedures for a priori classifications. Commun. Stat. Theor. Methods **5**, 1409–1424 (1976)
972. Mielke, P.W., Berry, K.J., Johnston, J.E.: Comparisons of continuous and discrete methods for combining probability values associated with matched-pairs t-test data. Percept. Motor Skill. **100**, 799–805 (2005)
973. Mielke, P.W., Berry, K.J., Johnston, J.E.: A FORTRAN program for computing the exact variance of weighted kappa. Percept. Motor Skill. **101**, 468–472 (2005)
974. Mielke, P.W., Berry, K.J., Johnston, J.E.: The exact variance of weighted kappa with multiple raters. Psychol. Rep. **101**, 655–660 (2007)
975. Mielke, P.W., Berry, K.J., Johnston, J.E.: Resampling programs for multiway contingency tables with fixed marginal frequency totals. Psychol. Rep. **101**, 18–24 (2007)

976. Mielke, P.W., Berry, K.J., Johnston, J.E.: Resampling probability values for weighted kappa with multiple raters. Psychol. Rep. **102**, 606–613 (2008)
977. Mielke, P.W., Berry, K.J., Johnston, J.E.: Unweighted and weighted kappa as measures of agreement for multiple judges. Int. J. Manag. **26**, 213–223 (2009)
978. Mielke, P.W., Berry, K.J., Johnston, J.E.: Robustness without rank order statistics. J. Appl. Stat. **38**, 207–214 (2011)
979. Mielke, P.W., Berry, K.J., Landsea, C.W., Gray, W.M.: Artificial skill and validation in meteorological forecasting. Weather Forecast. **11**, 153–169 (1996)
980. Mielke, P.W., Berry, K.J., Landsea, C.W., Gray, W.M.: A single-sample estimate of shrinkage in meteorological forecasting. Weather Forecast. **12**, 847–858 (1997)
981. Mielke, P.W., Berry, K.J., Medina, J.G.: Climax I and II: Distortion resistant residual analyses. J. Appl. Meterol. **21**, 788–792 (1982)
982. Mielke, P.W., Berry, K.J., Neidt, C.O.: A permutation test for multivariate matched-pairs analyses: Comparisons with Hotelling's multivariate matched-pairs T^2 test. Psychol. Rep. **78**, 1003–1008 (1996)
983. Mielke, P.W., Berry, K.J., Zelterman, D.: Fisher's exact test of mutual independence for $2 \times 2 \times 2$ cross-classification tables. Educ. Psychol. Meas. **54**, 110–114 (1994)
984. Mielke, P.W., Iyer, H.K.: Permutation techniques for analyzing multi-response data from randomized block experiments. Commun. Stat. Theor. Methods **11**, 1427–1437 (1982)
985. Mielke, P.W., Johnston, J.E., Berry, K.J.: Combining probability values from independent permutation tests: A discrete analog of Fisher's classical method. Psychol. Rep. **95**, 449–458 (2004)
986. Mielke, P.W., Long, M.A., Berry, K.J., Johnston, J.E.: g-treatment ridit analysis: Resampling permutation methods. Stat. Methodol. **6**, 223–229 (2009)
987. Mielke, P.W., Sen, P.K.: On asymptotic non-normal null distributions for locally most powerful rank test statistics. Commun. Stat. Theor. Methods **10**, 1079–1094 (1981)
988. Mielke, P.W., Siddiqui, M.M.: A combinatorial test for independence of dichotomous responses. J. Am. Stat. Assoc. **60**, 437–441 (1965)
989. Mielke, P.W., Yao, Y.C.: A class of multiple sample tests based on empirical coverages. Ann. Inst. Stat. Math. **40**, 165–178 (1988)
990. Mielke, P.W., Yao, Y.C.: On g-sample empirical coverage tests: Exact and simulated null distributions of test statistics with small and moderate sample sizes. J. Stat. Comput. Simul. **35**, 31–39 (1990)
991. Miettinen, O.S.: Comment on "Some reasons for not using the Yates continuity correction on 2×2 contingency tables" by W.J. Conover. J. Am. Stat. Assoc. **69**, 380–382 (1974)
992. Mifsud, C.J.: Algorithm 154: Combination in lexicographical order. Commun. ACM **6**, 103 (1963)
993. Mifsud, C.J.: Algorithm 155: Combination in any order. Commun. ACM **6**, 103 (1963)
994. Miller, R.G.: A trustworthy jackknife. Ann. Math. Stat. **35**, 1594–1605 (1964)
995. Miller, R.G.: The jacknife — a review. Biometrika **61**, 1–15 (1974)
996. Milton, R.C.: An extended table of critical values for the Mann–Whitney (Wilcoxon) two-sample statistic. J. Am. Stat. Assoc. **59**, 925–934 (1964)
997. Minkowski, H.: Über die positiven quadratishen formen und über kettenbruchähnliche algorithmen. Crelles J. **107**, 278–297 (1891)
998. Montgomery, D.C.: A conversation with Stu Hunter. Qual. Eng. **21**, 233–240 (2009)
999. Mood, A.M.: The distribution theory of runs. Ann. Math. Stat. **11**, 367–392 (1940)
1000. Mood, A.M.: Introduction to the Theory of Statistics. McGraw-Hill, New York (1950)
1001. Mood, A.M.: On the asymptotic efficiency of certain nonparametric two-sample tests. Ann. Math. Stat. **25**, 514–522 (1954)
1002. Moore, G.E.: Cramming more components onto integrated circuits. Electronics **38**, 114–117 (19 April 1965)
1003. Moran, P.A.P.: On the method of paired comparisons. Biometrika **34**, 363–365 (1947)
1004. Moran, P.A.P.: Rank correlation and permutation distributions. Math. Proc. Camb. **44**, 142–144 (1948)

1005. Moran, P.A.P.: Recent developments in ranking theory. J. R. Stat. Soc. B Met. **12**, 152–162 (1950)

1006. Mordkoff, J.T.: The assumption(s) of normality. http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%20I/I.07%20normal.pdf (2011). Accessed 18 Aug 2013

1007. Morrison, P., Morrison, P.: 100 or so books that shaped a century of science. Am. Sci. **87**, 542–544, 546, 549–550 (November–December, 1999) [This is an extensive book review of 100 books in the last century by Philip and Phylis Morrison for the *American Scientist*, published by Sigma Xi]

1008. Morton, N.: Cedric Smith (1917–2002). Int. Stat. Inst. News **26**, 9–10 (2002)

1009. Moscovici, S.: Obituary: Leon Festinger. Eur. J. Soc. Psychol. **19**, 263–269 (1989)

1010. Moses, L.E.: Non-parametric statistics for psychological research. Psychol. Bull. **49**, 122–143 (1952)

1011. Moses, L.E.: Statistical theory and research design. Annu. Rev. Psychol. **7**, 233–258 (1956)

1012. Mosteller, F.: Samuel S. Wilks: Statesman of statistics. Am. Stat. **18**, 11–17 (1964)

1013. Mundry, R.: Testing related samples with missing values: A permutation approach. Anim. Behav. **58**, 1143–1153 (1999)

1014. Munro, T.A.: Phenylketonuria: Data on forty-seven British families. Ann. Hum. Genet. **14**, 60–88 (1947)

1015. Murphy, K.R.: The passing of giants: Raymond B. Cattell and Jacob Cohen. Ind. Org. Psychol. http://www.siop.org/tip/backissues/tipapril98/obituary.aspx (April 1998). Accessed 21 July 2013

1016. Murphy, K.R., Cleveland, J.: Understanding Performance Appraisal: Social, Organizational, and Goal-based Perspectives. Sage, Thousand Oaks (1995)

1017. Murray, G.D.: Reply from *BJS* statistical advisor to “The Wilcoxon–Mann–Whitney test condemned” by J. Ludbrook. Br. J. Surg. **83**, 137 (1996)

1018. Murray, M.A.M.: The first lady of math? Yale Alum. Mag., 5–6. http://www.yalealumnimagazine.com/issues/2010_05/letters_412.html (May/June 2010). Accessed 14 Mar 2012

1019. Nadkarni, A.S.: Anant Raoji Kamat. Econ. Polit. Weekly **18**, 1351–1352 (30 July 1983)

1020. Nanda, D.N.: Distribution of the sum of roots of a determinantal equation. Ann. Math. Stat. **21**, 432–439 (1950)

1021. Narins, B. (ed.): World of Computer Science, vol. I. Thomson Gale, Farmington Hills (2002)

1022. Narula, S.C., Wellington, J.F.: An algorithm for the minimum sum of weighted absolute errors regression. Commun. Stat. Simul. C **B6**, 341–352 (1977)

1023. Narula, S.C., Wellington, J.F.: Prediction, linear regression and the minimum sum of relative errors. Technometrics **19**, 185–190 (1977)

1024. Narula, S.C., Wellington, J.F.: Selection of variables in linear regression using the minimum sum of weighted absolute errors criterion. Technometrics **21**, 299–306 (1979)

1025. Narula, S.C., Wellington, J.F.: The minimum sum of absolute errors regression: A state of the art survey. Int. Stat. Rev. **50**, 317–326 (1982)

1026. Neel, H.B.: Alvan R Feinstein, MD. Otolaryng. Head Neck **125**, 16 (2001)

1027. Nelder, J.A.: Present position and potential developments: Some personal views (with discussion). J. R. Stat. Soc. A Gen. **147**, 151–160 (1984)

1028. Nelder, J.A.: Yates, Frank. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 347–349. Wiley, New York (1997)

1029. Nelson, J.C., Pepe, M.S.: Statistical description of interrater variability in ordinal ratings. Stat. Methods Med. Res. **9**, 475–496 (2000)

1030. Neuhauser, D., Diaz, M.: Shuffle the deck, flip that coin: Randomization comes to medicine. Qual. Saf. Health Care **13**, 315–316 (2004)

1031. Neuhäuser, M.: The choice of α for one-sided tests. Drug Inf. J. **38**, 57–60 (2004)

1032. Newcomb, S.: Researches on the motion of the moon, Part II. The mean motion of the moon and other astronomical elements derived from observations of eclipses and occultations extending from the period of the Babylonians until A.D. 1908. Astron. Pap. **9**, 1–249 (1912)

1033. Neyman, J.: Statistical problems in agricultural experimentation (with discussion). Suppl. J. R. Stat. Soc. **B2**, 107–180 (1935) [As noted by Neyman: "With the cooperation of K. Iwaszkiewicz and St. Kołodziejczyk"]

1034. Neyman, J.: Mr. W. S. Gosset. J. Am. Stat. Assoc. **33**, 226–228 (1938)

1035. Neyman, J., Pearson, E.S.: On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika **20A**, 175–240 (1928)

1036. Neyman, J., Pearson, E.S.: On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. Biometrika **20A**, 263–294 (1928)

1037. Niederreiter, H.: Quasi-Monte Carlo methods and pseudorandom numbers. B. Math. Soc. **84**, 957–1041 (1978)

1038. Noether, G.E.: Asymptotic properties of the Wald–Wolfowitz test of randomness. Ann. Math. Stat. **21**, 231–246 (1950)

1039. Noether, G.E.: On a theorem of Pitman. Ann. Math. Stat. **26**, 64–68 (1955)

1040. Noether, G.E.: Comment on "Rank transformations as a bridge between parametric and nonparametric statistics" by W.J. Conover and R.L. Iman. Am. Stat. **35**, 129–130 (1981)

1041. Noreen, E.W.: Computer-intensive Methods For Testing Hypotheses: An Introduction. Wiley, New York (1989)

1042. Norman, R.: Biographies of women mathematicians: Grace Murray Hopper. http://www.agnesscott.edu/lriddle/women/hopper.htm (2001). Accessed 13 Mar 2012

1043. Norman, R.G., Scott, M.A.: Measurement of inter-rater agreement for transient events using Monte Carlo sampled permutations. Stat. Med. **26**, 931–942 (2007)

1044. O'Boyle, Jr., E., Aguinis, H.: The best and the rest: Revisiting the norm of normality of individual performance. Pers. Psychol. **65**, 79–119 (2012)

1045. O'Connor, A.: Dr. Alvan Feinstein, 75, innovator in diagnoses, dies. NY Times. http://www.nytimes.com/2001/10/29/nyregion/dr-alvan-feinstein-75-innovator-in-diagnoses-dies.html (29 October 2001). Accessed 9 Apr 2012

1046. O'Connor, J.J., Robertson, E.F.: James Stirling. http://www-history.mcs.st-and.ac.uk/Biographies/Stirling.html (1998). Accessed 16 Mar 2012

1047. O'Connor, J.J., Robertson, E.F.: Thomas Harriot. http://www-history.mcs.st-andrews.ac.uk/Biographies/Harriot.html (2000). Accessed 8 Sept 2012

1048. Odeh, R.E.: The generalized Mann–Whitney U-statistic. J. R. Stat. Soc. C Appl. **21**, 348–351 (1972)

1049. Odén, A., Wedel, H.: Arguments for Fisher's permutation test. Ann. Stat. **3**, 518–520 (1975)

1050. O'Gorman, T.W.: The performance of randomization tests that use permutations of independent variables. Commun. Stat. Simul. C **34**, 895–908 (2005)

1051. Ohashi, Y.: Randomization in cancer clinical trials: Permutation test and development of a computer program. Environ. Health Perspect. **87**, 13–17 (1990)

1052. Oja, H.: On permutation tests in multiple regression and analysis of covariance problems. Aust. J. Stat. **29**, 91–100 (1987)

1053. Okamoto, D.: Letter to the editor: Does it work for coffee? Significance **10**, 45–46 (June 2013)

1054. Olds, E.G.: Distribution of sums of squares of rank differences for small numbers of individuals. Ann. Math. Stat. **9**, 133–148 (1938)

1055. Olds, E.G.: The 5% significance levels for sums of squares of rank differences and a correction. Ann. Math. Stat. **20**, 117–118 (1949)

1056. Olkin, I.: A conversation with W. Allen Wallis. Stat. Sci. **6**, 121–140 (1991)

1057. Olkin, I.: A conversation with Churchill Eisenhart. Stat. Sci. **7**, 512–530 (1992)

1058. Olkin, I., Sampson, A.R.: Harold Hotelling. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 454–457. Springer, New York (2001)

1059. Olmstead, P.S., Tukey, J.W.: A corner test for association. Ann. Math. Stat. **18**, 495–513 (1947)

1060. Olson, J.: Henry Berthold Mann. Department of Mathematics, The Ohio State University. http://www.math.osu.edu/history/biographies/mann/. Accessed 20 Jan 2012

1061. Önder, H.: Using permutation tests to reduce type I and II errors for small ruminant research. J. Appl. Anim. Res. **32**, 69–72 (2007)
1062. Önder, H.: A comparative study of permutation tests with Euclidean and Bray–Curtis distances for common agricultural distributions in regression. J. Appl. Anim. Res. **34**, 133–136 (2008)
1063. Onghena, P., May, R.B.: Pitfalls in computing and interpreting randomization test p values: A commentary on Chen and Dunlap. Behav. Res. Methods Instrum. C **27**, 408–411 (1995)
1064. Ord, K.: In memoriam: Maurice George Kendall, 1907–1983. Am. Stat. **38**, 36–37 (1984)
1065. Ord-Smith, R.J.: Algorithm 308: Generation of permutations in pseudo-lexicographic order. Commun. ACM **10**, 7 (1967)
1066. Ord-Smith, R.J.: Remark on Langdon's algorithm. Commun. ACM **10**, 684 (1967)
1067. Ord-Smith, R.J.: Algorithm 323: Generation of permutations in lexicographic order. Commun. ACM **11**, 2 (1968)
1068. Ord-Smith, R.J.: Generation of permutation sequences, Part 1. Comput. J. **13**, 152–155 (1970)
1069. Ord-Smith, R.J.: Generation of permutation sequences, Part 2. Comput. J. **14**, 136–139 (1971)
1070. O'Reilly, F.J., Mielke, P.W.: Asymptotic normality of MRPP statistics from invariance principles of U-statistics. Commun. Stat. Theor. Methods **9**, 629–637 (1980)
1071. Orlowski, L.A., Grundy, W.D., Mielke, P.W., Schumm, S.A.: Geological applications of multi-response permutation procedures. Math. Geol. **25**, 483–500 (1993)
1072. Orlowski, L.A., Schumm, S.A., Mielke, P.W.: Reach classifications of the lower Mississippi river. Geomorphology **14**, 221–234 (1995)
1073. Orwell, G.: 1984. Harcourt Brace, New York (1949)
1074. Overbye, D.: Kenneth I. Appel, 80 dies; computerized higher math. NY Times **162**, A19 (29 April 2013)
1075. Owen, D.B.: Handbook of Statistical Tables. Addison-Wesley, Reading (1962)
1076. Pabst, M.R.: 1931 shifts academic tassels in sixty-sixth commencement. Vassar Misc. News **15**, 1–8 (10 June 1931)
1077. Pabst, M.R.: Properties of Bilinear Transformations in Unimodular Form. Department of Mathematics, University of Chicago, Chicago (1932) [This was Margaret Pabst's Master's thesis at the University of Chicago]
1078. Pabst, M.R.: The Public Welfare Administration of Dutchess County, New York. The Women's City and Country Club and Vassar College, Poughkeepsie (1933)
1079. Pabst, M.R.: Agricultural Trends in the Connecticut Valley Region of Massachusetts, 1800–1900, vol. 26. Smith College Studies in History, Northampton (1941)
1080. Pabst, W.R.: Notes: Appointments and resignations. Am. Econ. Rev. **28**, 865–878 (1938)
1081. Pagano, M., Taylor Halvorsen, K.: An algorithm for finding the exact significance levels of $r \times c$ contingency tables. J. Am. Stat. Assoc. **76**, 931–934 (1981)
1082. Pagano, M., Tritchler, D.: Algorithms for the analysis of several 2×2 contingency tables. SIAM J. Sci. Stat. Comput. **4**, 302–309 (1983)
1083. Pagano, M., Tritchler, D.: On obtaining permutation distributions in polynomial time. J. Am. Stat. Assoc. **78**, 435–440 (1983)
1084. Larry Page biography. http://www.biography.com/people/larry-page-12103347 (2012). Accessed 3 Nov 2012
1085. Page, E.S.: A note on generating random permutations. J. R. Stat. Soc. C Appl. **16**, 273–274 (1967)
1086. Pardo-Iguzquiza, E., Rodriguez-Tovar, F.J.: The permutation test as a non-parametric method for testing the statistical significance of power spectrum estimation in cyclostratigraphic research. Earth Planet. Sc. Lett. **181**, 175–189 (2000)
1087. Pareto, V.F.D.: L'economie et la sociologie au point de vue scientifique (Economics and sociology from a scientific perspective). In: Écrites Sociologiques Mineurs, vol. 22 of Oeuvres Complètes. Droz, Geneva (1980)

1088. Pascal, B.: Traité du triangle arithmétique (Treatise on the arithmetical triangle). In: Smith, D.E. (ed.) A Source Book in Mathematics, vol. I, pp. 67–79. Dover, New York (1959) [Translated by A. Savitsky]

1089. Patefield, W.M.: Algorithm 159: An efficient method of generating random $r \times c$ tables with given row and column totals. J. R. Stat. Soc. C Appl. **30**, 91–97 (1981)

1090. Patil, K.D.: Cochran's Q test: Exact distribution. J. Am. Stat. Assoc. **70**, 186–189 (1975)

1091. Payne, W.H., Ives, F.M.: Combination generators. ACM Trans. Math. Software **5**, 163–172 (1979)

1092. Pearson, E.S.: Untitled. Nature **124**, 615 (19 October 1929)

1093. Pearson, E.S.: Some aspects of the problem of randomization. Biometrika **29**, 53–64 (1937)

1094. Pearson, E.S.: 'Student' as a statistician. Biometrika **30**, 210–250 (1939)

1095. Pearson, E.S.: The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table. Biometrika **34**, 139–167 (1947)

1096. Pearson, E.S.: On questions raised by the combination of tests based on discontinuous distributions. Biometrika **37**, 383–398 (1950)

1097. Pearson, E.S.: The Neyman–Pearson story. In: David, F.N. (ed.) Research Papers in Statistics: Festschrift for J. Neyman. Wiley, London (1966)

1098. Pearson, E.S. (ed.): The History of Statistics in the 17th and 18th Centuries Against the Changing Background of Intellectual, Scientific and Religious Thought. Griffin, London (1978) [Lectures by Karl Pearson given at University College, London, during the academic sessions 1921–1933]

1099. Pearson, E.S.: Untitled. Nature **123** (8 June 1929) [Review by E.S. Pearson of the second edition of R.A. Fisher's *Statistical Methods for Research Workers*]

1100. Pearson, E.S., Adyanthāya, N.K.: The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. Biometrika **21**, 259–286 (1929)

1101. Pearson, E.S., Hartley, H.O. (eds.): Biometrika Tables for Statisticians, vol. I. Cambridge University Press, Cambridge (1954)

1102. Pearson, E.S., Hartley, H.O. (eds.): Biometrika Tables for Statisticians, vol. II. Cambridge University Press, Cambridge (1972)

1103. Pearson, E.S., Kendall, M.G. (eds.): Studies in the History of Statistics and Probability, vol. I. Griffin, London (1970)

1104. Pearson, K.: Contributions to the mathematical theory of evolution. Proc. R. Soc. Lond. **54**, 329–333 (1893)

1105. Pearson, K.: Contributions to the mathematical theory of evolution. Philos. Trans. R. Soc. Lond. A **185**, 71–110 (1894)

1106. Pearson, K.: Contributions to the mathematical theory of evolution, II. Skew variation in homogeneous material. Philos. Trans R. Soc. Lond. A **186**, 343–414 (1895)

1107. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. 5 **50**, 157–175 (1900)

1108. Pearson, K.: On the laws of inheritance in man: II. On the inheritance of the mental and moral characters in man, and its comparison with the inheritance of the physical characters. Biometrika **3**, 131–190 (1904)

1109. Pearson, K.: Mathematical contributions to the theory of evolution, XVI. On further methods of determining correlation. In: Drapers' Company Research Memoirs, Biometric Series IV, pp. 1–39. Dulau and Company, London (1907)

1110. Pearson, K.: On the coefficient of racial likeness. Biometrika **18**, 105–117 (1926)

1111. Peatman, J.G., Shafer, R.: A table of random numbers from Selective Service numbers. J. Psychol. **14**, 295–305 (1942)

1112. Peck, J.E.L., Schrack, G.F.: Algorithm 86: PERMUTE. Commun. ACM **5**, 208–209 (1962)

1113. Peirce, C.S., Jastrow, J.: On small differences in sensation. Natl. Acad. Sci. Bio. Mem. **3**, 73–83 (1885) [Publication of an address given to the National Academy of Sciences on 17 October 1884]

1114. Pellicane, P.J., Mielke, P.W.: Median-based regression methods in wood science applications. Wood Sci. Technol. **27**, 249–256 (1993)

1115. Pellicane, P.J., Mielke, P.W.: Permutation procedures for multi-dimensional applications in wood related research. Wood Sci. Technol. **33**, 1–13 (1999)

1116. Pellicane, P.J., Potter, R.S., Mielke, P.W.: Permutation procedures as a statistical tool in wood related applications. Wood Sci. Technol. **23**, 193–204 (1989)

1117. Peritz, E.: Comment on “Yates’s correction for continuity and the analysis of 2 × 2 contingency tables” by M.G. Haviland. Stat. Med. **11**, 845 (1992)

1118. Perkins, S.M., Becker, M.P.: Assessing rater agreement using marginal association models. Stat. Med. **21**, 1743–1760 (2002)

1119. Perks, J.: Commentary: ‘The next trick is impossible’. Int. J. Epidemiol. **39**, 1153–1155 (2010)

1120. Pesarin, F.: Multivariate Permutation Tests: With Applications in Biostatistics. Wiley, Chichester (2001)

1121. Pesarin, F., Salmaso, L.: Exact permutation tests for unreplicated factorials. Appl. Stoch. Model. Bus. **18**, 287–299 (2002)

1122. Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data: Theory, Applications and Software. Wiley, Chichester (2010)

1123. Pfaffenberger, R., Dinkel, J.: Absolute deviations curve-fitting: An alternative to least squares. In: David, H.A. (ed.) Contributions to Survey Sampling and Applied Statistics, pp. 279–294. Academic, New York (1978)

1124. Phillips, J.P.N.: Algorithm 28: Permutations of the elements of a vector in lexicographic order. Comput. J. **10**, 311 (1967)

1125. Phillips, J.P.N.: A simplified accurate algorithm for the Fisher–Yates exact test. Psychometrika **47**, 349–351 (1982)

1126. Picard, R.: Randomization and design: II. In: Feinberg, S.E., Hinkley, D.V. (eds.) R. A. Fisher: An Appreciation, pp. 46–58. Springer, Heidelberg (1980)

1127. Pierce, A.: Fundamentals of Nonparametric Statistics. Dickenson, Belmont (1970)

1128. Pillai, K.C.S.: Some new test criteria in multivariate analysis. Ann. Math. Stat. **26**, 117–121 (1955)

1129. Pitman, E.J.G.: Significance tests which may be applied to samples from any populations. Suppl. J. R. Stat. Soc. **4**, 119–130 (1937)

1130. Pitman, E.J.G.: Significance tests which may be applied to samples from any populations: II. The correlation coefficient test. Suppl. J. R. Stat. Soc. **4**, 225–232 (1937)

1131. Pitman, E.J.G.: Significance tests which may be applied to samples from any populations: III. The analysis of variance test. Biometrika **29**, 322–335 (1938)

1132. Pitman, E.J.G.: Lecture notes on non-parametric statistical inference (1948) [Unpublished lecture notes for a course given at Columbia University in 1948]

1133. Pitman, E.J.G.: Reminiscences of a mathematician who strayed into statistics. In: Gani, J. (ed.) The Making of Statisticians, pp. 111–125. Springer, New York (1982)

1134. Pitt, D.G., Kreutzweiser, D.P.: Applications of computer-intensive statistical methods to environmental research. Ecotox. Environ. Saf. **39**, 78–97 (1998)

1135. Plackett, R.L.: The continuity correction in 2 × 2 tables. Biometrika **51**, 327–337 (1964)

1136. Plackett, R.L.: Random permutations. J. R. Stat. Soc. B Met. **30**, 517–534 (1968)

1137. Plackett, R.L.: The marginal totals of a 2 × 2 table. Biometrika **64**, 37–42 (1977)

1138. Plackett, R.L.: The Analysis of Categorical Data, 2nd edn. Macmillan, New York (1981)

1139. Plackett, R.L.: Discussion of “Tests of significance in 2 × 2 tables” by F. Yates. J. R. Stat. Soc. A Gen. **147**, 458 (1984)

1140. Plackett, R.L.: Obituary: Churchill Eisenhart. J. R. Stat. Soc. A Stat. **158**, 338 (1995)

1141. Porter, T.M.: The Rise of Statistical Thinking, 1820–1900. Princeton University Press, Princeton (1986)

1142. Portnoy, S., Koenker, R.: The Gausssian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators (with discussion). Stat. Sci. **12**, 279–300 (1997)

1143. Potvin, C., Roff, D.A.: Distribution-free and robust statistical methods: Viable alternatives to parametric statistics? Ecology **74**, 1617–1628 (1993)
1144. Powell, M.: A redoubt of learning holds firm: The Royal Society, crucible of the scientific revolution that formed the modern world, strives to stay relevant. NY Times **162**, D1–D2 (4 September 2012)
1145. Quenouille, M.H.: Approximate tests of correlation in time-series. J. R. Stat. Soc. B Met. **11**, 68–84 (1949)
1146. Quenouille, M.H.: Notes on bias in estimation. Biometrika **43**, 353–360 (1956)
1147. Quetelet, L.A.J.: Lettres à S. A. R. le Duc Régnant de Saxe-Cobourg et Gotha, sur la Théorie des Probabilitiés Appliquée aux Sciences Morales et Politiques. Hayez, Bruxelles (1846) [English translation, *Letters Addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha on the Theory of Probabilities as Applied to the Moral and Political Sciences*, by O.G. Downes and published by Charles & Edwin Layton, London, 1849]
1148. Raab, G.M., Butcher, I.: Randomization inference for balanced cluster-randomized trials. Clin. Trials **2**, 130–140 (2005)
1149. Rabinowitz, M., Berenson, M.L.: A comparison of various methods of obtaining random order statistics for Monte Carlo computations. Am. Stat. **28**, 27–29 (1974)
1150. Radlow, R., Alf, Jr., E.F.: An alternate multinomial assessment of the accuracy of the χ^2 test of goodness of fit. J. Am. Stat. Assoc. **70**, 811–813 (1975)
1151. Ralph, N.: Processors: What to expect from CPUs in 2012. http://www.pcworld.com/article/246688/processors_what_to_expect_from_cpus_in_2012.html (27 December 2011). Accessed 29 Apr 2012
1152. RAND Corporation: A Million Random Digits with 100,000 Normal Deviates. Glencoe Free Press, Glencoe (1955)
1153. Randles, R.H., Wolfe, D.A.: Introduction to the Theory of Nonparametric Statistics. Wiley, New York (1979)
1154. Rao, C.R.: Generation of random permutations of given number of elements using random sampling numbers. Sankhyā **23**, 305–307 (1961)
1155. Rao, C.R.: R. A. Fisher: The founder of modern statistics. Stat. Sci. **7**, 34–48 (1992)
1156. Rao, J.S., Murthy, V.K.: A two-sample nonparametric test based on spacing frequencies. In: Proceeding of the 43rd Session of the International Statistical Institute, vol. 43, pp. 223–227. International Statistical Institute, The Hague (1981)
1157. Read, T.R.C., Cressie, N.A.C.: Goodness-of-Fit for Discrete Multivariate Data. Springer, New York (1988)
1158. Redin, J.: A brief history of mechanical calculators: Part III, Getting ready for the 20th century. http://www.xnumber.com/xnumber/mechanical3.htm. Accessed 12 Mar 2012
1159. Reich, R.M., Mielke, P.W., Hawksworth, F.G.: Spatial analysis of ponderosa pine trees infected with dwarf mistletoe. Can. J. Forest Res. **21**, 1808–1815 (1991)
1160. Reid, C.: Neyman — From Life. Springer, New York (1982)
1161. Reid, N.: The roles of conditioning in inference. Stat. Sci. **10**, 138–157 (1995) [See also the accompanying discussion in the same issue on pages 173–199]
1162. Reilly, E.D.: Milestones in Computer Science and Information Technology. Greenwood Press, Westport (2003)
1163. Reiss, P.T., Stevens, M.H.H., Shehzad, Z., Petkova, E., Milham, M.P.: On distance-based permutation tests for between-group comparisons. Biometrics **66**, 636–643 (2010)
1164. Reyment, R.A.: An idiosyncratic history of early morphometrics. In: Marcus, L.F., Corti, M., Lay, A., Naylor, G.J.P., E, S.D. (eds.) Advances in Morphometrics, vol. 284, Life Sciences, pp. 15–22. Plenum Press, New York (1996) [Proceedings of the NATO Advanced Study Institute on Advances in Morphometrics, held at Il Ciocco, Tuscany, Italy in July 1993]
1165. Rhodes, E.C.: Reducing observations by the method of minimum deviations. Philos. Mag. 7 **9**, 974–992 (1930)
1166. Rhodes, R.: Hedy's Folly: The Life and Breakthrough Inventions of Hedy Lamarr, the Most Beautiful Woman in the World. Doubleday, New York (2011)

1167. Rice, W.R.: A new probability model for determining exact p-values for 2×2 contingency tables when comparing binomial proportions. Biometrics **44**, 1–22 (1988)
1168. Rice, W.R.: Reply to the discussion of "A new probability model for determining exact P-values for 2×2 contingency tables when comparing binomial proportions" by W.R. Rice. Biometrics **44**, 18–22 (1988)
1169. Richards, L.E., Byrd, J.: Algorithm 304: Fisher's randomization test for two small independent samples. J. R. Stat. Soc. C Appl. **45**, 394–398 (1996)
1170. Richardson, J.T.E.: Variants of chi-square for 2×2 contingency tables. Br. J. Math. Stat. Psychol. **43**, 309–326 (1990)
1171. Richardson, J.T.E.: The analysis of 2×1 and 2×2 contingency tables: A historical review. Stat. Methods Med. Res. **3**, 107–134 (1994)
1172. Ripley, B.D.: Stochastic Simulation. Wiley, New York (1987)
1173. Robbins, H.: A remark on Stirling's formula. Am. Math. Mon. **62**, 26–29 (1955)
1174. Robertson, W.H.: Programming Fisher's exact method of comparing two percentages. Technometrics **2**, 103–107 (1960)
1175. Robinson, A.: On the shoulders of giants. In: Robinson, A. (ed.) The Scientists: An Epic of Discovery, pp. 7–15. Thames & Hudson, London (2012)
1176. Robinson, A.P., Hamann, J.D.: Forest Analytics with *R*. Springer, New York (2011)
1177. Robinson, C.L.: Algorithm 317: PERMUTATION. Commun. ACM **10**, 11 (1967)
1178. Robinson, J.: The large-sample power of permutation tests for randomization models. Ann. Stat. **1**, 291–296 (1973)
1179. Robinson, J.: Approximations to some test statistics for permutation tests in a completely randomized design. Aust. J. Stat. **25**, 358–369 (1983)
1180. Rocke, A.: August Kekulé. In: Robinson, A. (ed.) The Scientists: An Epic of Discovery, pp. 133–135. Thames & Hudson, London (2012)
1181. Rodden, B.E.: In defense of Langdon's algorithm. Commun. ACM **11**, 150 (1968)
1182. Rodrigues, O.: Note sur les inversions, ou dérangements produits dans les permutations (Note on inversions, or products of derangements in permutations). J. Math. Pure. Appl. **4**, 236–240 (1839) [The Journal de Mathématiques Pures et Appliquées is also known as the Journal de Liouville]
1183. Rohl, J.S.: Programming improvements to Fike's algorithm for generating permutations. Comput. J. **19**, 156–159 (1976)
1184. Rohl, J.S.: Generating permutations by choosing. Comput. J. **21**, 302–305 (1978)
1185. Rohl, J.S.: Ord-Smith's pseudo-lexicographical permutation procedure is the Tompkins–Paige algorithm. Comput. J. **34**, 569–570 (1991)
1186. Röhmel, J.: The permutation distribution of the Friedman test. Comput. Stat. Data Anal. **26**, 83–99 (1997)
1187. Rojo, J.: Erich Leo Lehmann — A glimpse into his life and work. Ann. Stat. **39**, 2244–2265 (2011)
1188. Rolfe, T.J.: Randomized shuffling. Dr. Dobbs J. **25**, 113–114 (2000)
1189. Romano, J.P.: Bootstrap and randomization tests of some nonparametric hypotheses. Ann. Stat. **17**, 141–159 (1989)
1190. Romesburg, H.C.: Exploring, confirming, and randomization tests. Comput. Geosci. **11**, 19–37 (1985)
1191. Romesburg, H.C., Marshall, K., Mauk, T.P.: FITEST: A computer program for "exact chi-square" goodness-of-fit significance tests. Comput. Geosci. **7**, 47–58 (1981)
1192. Ronchetti, E.: Bounded influence inference in regression: A review. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-norm and Related Methods, pp. 65–80. Elsevier, Amsterdam (1987) [Collection of invited papers presented at The First International Conference on Statistical Data Analysis Based on the L_1-norm and Related Methods, held in Neuchâtel, Switzerland, from 31 August to 4 September 1987]
1193. Rosenbaum, S.: Tables for a nonparametric test of dispersion. Ann. Math. Stat. **24**, 663–668 (1953)

1194. Rosenberger, J.L., Gasko, M.: Comparing location estimators: Trimmed means, medians, and trimean. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (eds.) Understanding Robust and Exploratory Data Analysis, pp. 297–338. Wiley, New York (1983)

1195. Ross, G.: Sir John Russell (1872–1965). Harpenden History. http://www.harpenden-history.org.uk/page_id__203_path__0p3p.aspx (2011). Accessed 8 Nov 2012

1196. Ross, G.: Fisher and the Millionaire: The statistician and the calculator. Significance **9**, 46–48 (2012)

1197. Routledge, R.D.: Resolving the conflict over Fisher's exact test. Can. J. Stat. **20**, 201–209 (1992)

1198. Routledge, R.D.: P-values from permutation and F-tests. Comput. Stat. Data Anal. **24**, 379–386 (1997)

1199. Roy, M.K.: Evaluation of permutation algorithms. Comput. J. **21**, 296–301 (1978)

1200. Roy, T.: The effect of heteroscedasticity and outliers on the permutation t-test. J. Stat. Comput. Simul. **72**, 23–26 (2002)

1201. Royo, J., Ferrer, S.: Tables of random numbers obtained from numbers in the Spanish national lottery. Trabajos Estad. **5**, 247–256 (1954)

1202. Rubin, D.B.: Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. J. Am. Stat. Assoc. **75**, 591–593 (1980)

1203. Rubin, D.B.: Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat. Sci. **5**, 472–480 (1990)

1204. Rubinstein, R.Y.: Simulation and the Monte Carlo Methods. Wiley, New York (1981)

1205. Rucci, A.J., Tweney, R.D.: Analysis of variance and the "second discipline" of scientific psychology: A historical account. Psychol. Bull. **87**, 166–184 (1980)

1206. Rümke, C.L., van Eeden, C.: Statistiek voor Medici. Uitgeverij L. Stafleu and Zoon, Leiden (1961)

1207. Russell, C.M., Bradley, E.L., Retief, D.H.: A permutation test for regression analysis of dental research data. J. Dent. Res. **69**, 127–127 (1990)

1208. Russell, C.M., Martin, J.A.: Multivariate statistical analysis for radiographic cephalometry: A permutation test method. J. Dent. Res. **73**, 271–271 (1994)

1209. Russell, E.J.: Rothamsted and its experimental station. Agric. Hist. **4**, 161–183 (1942)

1210. Russell, E.J.: A History of Agricultural Science in Great Britain. Allen & Unwin, London (1966)

1211. Russell, E.R.: Rothamsted manor house. Records of the Rothamsted Staff, Harpenden **5**, 23–29 (June 1935). [Published by the Rothamsted Experimental Station]

1212. Saal, F.E., Downey, R.G., Lahey, M.A.: Rating the ratings: Assessing the quality of rating data. Psychol. Bull. **88**, 413–428 (1980)

1213. Sag, T.W.: Permutations of a set with repetitions. Commun. ACM **7**, 585 (1964)

1214. Saito, M., Matsumoto, M.: SIMD-oriented fast Mersenne twister: A 128-bit pseudorandom number generator. In: Keller, A., Heinrich, S., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2006, pp. 607–622. Springer, Berlin (2008) [Proceedings of the Seventh International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, held at Ulm University, Germany, in August 2006]

1215. Sakaori, F.: Permutation test for equality of correlation coefficients in two populations. Commun. Stat. Simul. C **31**, 641–651 (2002)

1216. Sakoda, J.M., Cohen, B.H.: Exact probabilities for contingency tables using binomial coefficients. Psychometrika **22**, 83–86 (1957)

1217. Salama, I.A., Quade, D.: A note on Spearman's footrule. Commun. Stat. Simul. C **19**, 591–601 (1990)

1218. Salsburg, D.: The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Holt, New York (2001)

1219. Sampson, A.R.: Harold Hotelling. StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies. http://statprob.com/encyclopedia/HaroldHOTELLING.html (2012). Accessed 29 Nov 2012

1220. Sandelius, M.: A simple randomization procedure. J. R. Stat. Soc. B Met. **24**, 472–481 (1962)

1221. Sandiford, P.: Educational Psychology. Longmans, Green & Company, New York (1928) [The graphical method appears in an Appendix by S.D. Holmes, 'A graphical method of estimating *R* for small groups', pp. 391–394]

1222. Sattolo, S.: An algorithm to generate a random cyclic permutation. Inform. Process. Lett. **22**, 315–317 (1986)

1223. Saunders, I.A.: Algorithm 205: Enumeration of $R \times C$ tables with repeated row totals. J. R. Stat. Soc. C Appl. **33**, 340–352 (1984)

1224. Savage, I.R.: Nonparametric statistics. J. Am. Stat. Assoc. **52**, 331–344 (1957)

1225. Savage, I.R.: Bibliography of Nonparametric Statistics. Harvard University Press, Cambridge (1962)

1226. Savage, L.J.: On rereading R. A. Fisher. Ann. Stat. **4**, 441–500 (1976)

1227. Sawrey, W.L.: A distinction between exact and approximate nonparametric methods. Psychometrika **23**, 171–177 (1958)

1228. Scattergood, B.P.: The manor house of Rothamsted and the Wittewronge descent of Sir John Bennet Lawes, F.R.S. Records of the Rothamsted Staff, Harpenden **4**, 21–32 (February 1933) [Published by the Rothamsted Experimental Station]

1229. Schachter, S.: Leon Festinger, May 8, 1919 – February 11, 1989. Natl. Acad. Sci. Bio. Mem., 99–110. http://www.motherjones.com/files/lfestinger.pdf (1994). Accessed 19 June 2012

1230. Scheffé, H.: Statistical inference in the non-parametric case. Ann. Math. Stat. **14**, 305–332 (1943)

1231. Scheffé, H.: Alternative models for the analysis of variance. Ann. Math. Stat. **27**, 251–271 (1956)

1232. Scheffé, H.: The Analysis of Variance. Wiley, New York (1959)

1233. Schmidt, F.L., Johnson, R.H.: Effect of race on peer ratings in an industrial situation. J. Appl. Psychol. **57**, 237–241 (1973)

1234. Schouten, H.J.A.: Measuring pairwise agreement among many observers. Biometrical J. **22**, 497–504 (1980)

1235. Schouten, H.J.A.: Measuring pairwise agreement among many observers: II. Some improvements and additions. Biometrical J. **24**, 431–435 (1982)

1236. Schouten, H.J.A.: Measuring pairwise interobserver agreement when all subjects are judged by the same observers. Stat. Neerl. **36**, 45–61 (1982)

1237. Schouten, H.J.A., Molenaar, I.W., van Strik, R., Boomsma, A.: Comparing two independent binomial proportions by a modified chi square test. Biometrical J. **22**, 241–248 (1980)

1238. Schrack, G.F., Shimrat, M.: Algorithm 102: Permutation in lexicographical order. Commun. ACM **5**, 346 (1962)

1239. Schuster, C.: A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. Educ. Psychol. Meas. **64**, 243–253 (2004)

1240. Schuster, C., Smith, D.A.: Dispersion-weighted kappa: An integrative framework for metric and nominal scale agreement coefficients. Psychometrika **70**, 135–146 (2005)

1241. Scott, E.L.: Neyman, Jerzy. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 137–145. Wiley, New York (1997)

1242. Sedgewick, R.: Permutation generation methods. Comput. Surv. **9**, 137–164 (1977)

1243. Seel, P.B.: Digital Universe: The Global Telecommunication Revolution. Wiley-Blackwell, West Sussex (2012)

1244. Seigel, D.G., Podgor, M.J., Remaley, N.A.: Acceptable values of kappa for comparison of two groups. Am. J. Epidemiol. **135**, 571–578 (1992)

1245. Sellers, W.C., Yeatman, R.J.: 1066 and All That: A Memorable History of England. Methuen, London (1930)

1246. Sen, P.K.: On some multisample permutation tests based on a class of *U*-statistics. J. Am. Stat. Assoc. **62**, 1201–1213 (1965)

1247. Sen, P.K.: On some permutation tests based on U-statistics. Calcutta Stat. Assoc. **14**, 106–126 (1965)
1248. Seneta, E.: The weighted median and multiple regression. Aust. J. Stat. **25**, 370–377 (1983)
1249. Seneta, E., Steiger, W.L.: A new LAD curve-fitting algorithm: Slightly overdetermined equation systems in l_1. Discrete Appl. Math. **7**, 79–91 (1984)
1250. Senn, S.: Fisher's game with the devil. Stat. Med. **13**, 217–230 (1994) [Publication of a paper presented at the Statisticians in the Pharmaceutical Industry (PSI) annual conference held in September 1991 in Bristol, England]
1251. Senn, S.: Tea for three: Of infusions and inferences and milk in first. Significance **9**, 30–33 (December 2012)
1252. Senn, S.: Response to "Tea break" by S. Springate. Significance **10**, 46 (June 2013)
1253. Shah, A.: Intel unveils new core processors code-named Ivy Bridge. http://www.itworld.com/hardware/270726/intel-unveils-new-core-processors-code-named-ivy-bridge (23 April 2012). Accessed 29 Apr 2012
1254. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
1255. Shao, X.M.: An efficient algorithm for the exact test on unordered $2 \times J$ contingency tables with equal column sums. Comput. Stat. Data Anal. **25**, 273–285 (1997)
1256. Shapiro, F.R. (ed.): The Yale Book of Quotations. Yale University Press, New Haven (2006)
1257. Sheehan, W., Westfall, J.: The Transits of Venus. Prometheus, Amherst (2004)
1258. Shen, M.K.: On the generation of permutations and combinations. BIT **2**, 228–231 (1962)
1259. Shen, M.K.: Algorithm 202: Generation of permutations in lexicographical order. Commun. ACM **6**, 517 (1963)
1260. Shepherd, J.: World education rankings: Which country does best at reading, maths and science?. http://www.guardian.co.uk/news/datablog/2010/dec/07/world-education-rankings-maths-science-reading (2010). Accessed 16 Feb 2012
1261. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, 3rd edn. Chapman & Hall/CRC, Boca Raton (2004)
1262. Shewhart, W.A., Winters, F.W.: Small samples — New experimental results. J. Am. Stat. Assoc. **23**, 144–153 (1928)
1263. Sheynin, O.: Theory of Probability: A Historical Essay, 2nd edn. Oscar Sheynin, Berlin (2009)
1264. Sheynin, O.B.: R. J. Boscovich's work on probability. Arch. Hist. Exact Sci. **9**, 306–324 (1973)
1265. Shields, E.T.: Electronic research tools: Thomas Harriot surfs the web. Thomas Harriot College of Arts and Sciences. http://www.ecu/cs-cas/harriot/thomasharriotontheweb.cfm (2008). Accessed 9 Sept 2012
1266. Shirley, J.W.: Binary numeration before Leibniz. Am. J. Phys. **19**, 452–454 (1951)
1267. Shirley, J.W.: Thomas Harriot: A Biography. Clarendon, Oxford (1983)
1268. Short, J.: An account of the transit of Venus over the Sun, on Saturday morning, 6th June 1761, at Savile-House, about $8''$ of time west of St. Paul's, London. Philos. Trans. R. Soc. Lond. **52**, 178–182 (1761–1762) [Published in the Philosophical Transactions of the Royal Society of London (1683–1775)]
1269. Short, J.: The observations of the internal contact of Venus with the Sun's limb, in the late transit, made in different places of Europe, compared with the time of the same contact observed at the Cape of Good Hope, and the parallax of the Sun from thence determined. By James Short, A.M. F.R.S. Philos. Trans. R. Soc. Lond. **52**, 611–628 (1761–1762) [Published in the Philosophical Transactions of the Royal Society of London (1683–1775)]
1270. Shrout, P.E., Spitzer, R.L., Fleiss, J.L.: Quantification of agreement in psychiatric diagnosis revisited. Arch. Gen. Psychiatr. **44**, 172–177 (1987)
1271. Siegel, A.E.: Sidney Siegel: A Memoir. In: Messick, S., Brayfield, A.H. (eds.) Decision and Choice: Contributions of Sidney Siegel, pp. 1–23. McGraw-Hill, New York (1964)
1272. Siegel, S.: Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York (1956)

1273. Siegel, S., Tukey, J.W.: A nonparametric sum of ranks procedure for relative spread in unpaired samples. J. Am. Stat. Assoc. **55**, 429–445 (1960) [Corrigendum: J. Am. Stat. Assoc. **56**, 1005 (1961)]

1274. Siegfried, T.: Odds are, it's wrong. Sci. News **177**, 26–29 (27 March 2010)

1275. Silvey, S.D.: The equivalence of asymptotic distributions under randomisation and normal theories. Proc. Glasgow Math. Assoc. **1**, 139–147 (1953)

1276. Silvey, S.D.: The asymptotic distributions of statistics arising in certain nonparametric tests. Proc. Glasgow Math. Assoc. **2**, 47–51 (1954)

1277. Simon, J.L.: Resampling: The New Statistics. Duxbury, Pacific Grove (1997)

1278. Singer, I.B.: The Manor and the Estate. University of Wisconsin Press, Madison (2004)

1279. Singleton, R.R.: A method for minimizing the sum of absolute values of deviations. Ann. Math. Stat. **11**, 301–310 (1940)

1280. Sketch of Sir John Bennet Lawes. http://www.popsci.com/archive-viewer?id=-CQDAAAAMBAJ&pg=null&query=john%20bennet%20lawes (March 1886). Accessed 20 July 2012

1281. Skipper, J.K., Guenther, A.L., Nass, G.: The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. Am. Sociol. **2**, 16–18 (1967)

1282. Smalley, E.: Ultimate Computing. Discover, pp. 10–11 (July/August 2011)

1283. Smirnov, N.V.: On the estimation of discrepancy between empirical curves of distribution for two independent samples. Bull. Math. Univ. Moscov **2**, 3–16 (1939)

1284. Smith, J.Y.: William R. Pabst dies at 80; Navy quality control official. Wash. Post **105**, C4 (1992)

1285. Smith, P.L., Johnson, L.R., Priegnitz, D.L., Boe, B.A., Mielke, P.W.: An exploratory analysis of crop hail insurance data for evidence of cloud seeding effect in North Dakota. J. Appl. Meterol. **36**, 463–473 (1997)

1286. Smith, W.B.: Herman Otto Hartley (1912–1980). Am. Stat. **35**, 142–143 (1981)

1287. Smith, W.B.: H.O. Hartley (1912–1980) Revered and remembered. Amstat News **339**, 10–11 (September 2005)

1288. Smith, W.L.: Harold Hotelling. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 123–124. Wiley, New York (1997)

1289. Snedecor, G.W.: Calculation and Interpretation of Analysis of Variance and Covariance. Collegiate Press, Ames (1934)

1290. Snyder, L.J.: The Philosophical Breakfast Club: Four Remarkable Friends Who Transformed Science and Changed the World. Random House, New York (2011)

1291. Soeken, K.L., Prescott, P.A.: Issues in the use of kappa to estimate reliability. Med. Care **24**, 733–741 (1986)

1292. Sohn, D.: Knowledge in psychological science: That of process or of population? J. Psychol. **126**, 5–16 (1992)

1293. Solow, A.R.: A randomization test for misclassification probability in discriminant analysis. Ecology **7**, 2379–2382 (1990)

1294. Somers, R.H.: A new asymmetric measure of association for ordinal variables. Am. Sociol. Rev. **27**, 799–811 (1962)

1295. Somers, R.H.: A similarity between Goodman and Kruskal's tau and Kendall's tau, with a partial interpretation of the latter. J. Am. Stat. Assoc. **57**, 804–812 (1962)

1296. Soms, A.P.: An algorithm for the discrete Fisher's permutation test. J. Am. Stat. Assoc. **72**, 662–664 (1977)

1297. Soper, H.E., Young, A.W., Cave, B.M., Lee, A., Pearson, K.: On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. Biometrika **11**, 328–413 (1916)

1298. Sowey, E.R.: A chronological and classified bibliography on random number generation and testing. Int. Stat. Rev. **40**, 355–371 (1972)

1299. Sparkes, M.: MareNostrum, the world's most gorgeous super-computer. http://gizmodo.com/293608/marenostrum-the-worlds-most-gorgeous-super+computer (2007). Accessed 12 Mar 2012

1300. Spearman, C.E.: The proof and measurement of association between two things. Am. J. Psychol. **15**, 72–101 (1904)
1301. Spearman, C.E.: ‘Footrule’ for measuring correlation. Br. J. Psychol. **2**, 89–108 (1906)
1302. Spearman, C.E.: Correlation calculated from faulty data. Br. J. Psychol. **3**, 271–295 (1910)
1303. Spearman, C.E.: C. Spearman. In: Murchison, C. (ed.) A History of Psychology in Autobiography, vol. I, pp. 299–333. Russell and Russell, New York (1961)
1304. Speed, T.P.: Introductory remarks on Neyman (1923). Stat. Sci. **5**, 463–464 (1990)
1305. Spenser, J.E.: Robert Charles Geary. StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies. http://statprob.com/encyclopedia/RobertCharlesGEARY.html. Accessed 20 Dec 2011
1306. Spenser, J.E.: Robert Charles Geary 1896–1983. Econometrica **51**, 1599–1601 (1983)
1307. Spenser, J.E.: Robert Charles Geary. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 459–463. Springer, New York (2001)
1308. Spino, C., Pagano, M.: Efficient calculation of the permutation distribution of robust two-sample statistics. Comput. Stat. Data Anal. **12**, 349–365 (1991)
1309. Spino, C., Pagano, M.: Efficient calculation of the permutation distribution of trimmed means. J. Am. Stat. Assoc. **86**, 729–737 (1991)
1310. Spitzer, W.O.: The teacher’s teacher: A personal tribute to Alvan R. Feinstein. J. Epidemiol. Commun. H. **56**, 328–329 (2002)
1311. Spitznagel, E.L., Helzer, J.E.: A proposed solution to the base rate problem in the kappa statistic. Arch. Gen. Psychiatr. **42**, 725–728 (1985)
1312. Spława-Neyman, J.: Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych (On the application of probability theory to agricultural experiments. Essay on principles. Section 9). Rocz. Nauk Rolnicz. (Ann. Agric. Sci.) **10**, 1–51 (1923) [Translated from the original Polish by D. M. Dabrowska and T. P. Speed and published in Stat. Sci. **5**, 465–472 (1990)]
1313. Springate, S.: Tea break. Significance **10**, 45–46 (February 2013)
1314. Sprott, D.A.: A note on a class of occupancy problems. Am. Stat. **23**, 12–13 (1969)
1315. Starmer, C.F., Grizzle, J.E., Sen, P.K.: Comment on “Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables” by W.J. Conover. J. Am. Stat. Assoc., 376–378 (1974)
1316. Steinhaus, H.: Table of shuffled fourdigit numbers. Rozp. Matemat. **6**, 1–46 (1954)
1317. Stigler, S.M.: Studies in the history of probability and statistics, XXXII. Laplace, Fisher, and the discovery of the concept of sufficiency. Biometrika **60**, 439–445 (1973)
1318. Stigler, S.M.: American Contributions to Mathematical Statistics in the Nineteenth Century, vol. I. Arno Press, New York (1980)
1319. Stigler, S.M.: American Contributions to Mathematical Statistics in the Nineteenth Century, vol. II. Arno Press, New York (1980)
1320. Stigler, S.M.: The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge (1986)
1321. Stigler, S.M.: Statistics on the Table: The History of Statistical Concepts and Methods. Harvard University Press, Cambridge (1999)
1322. Stigler, S.M.: Fisher in 1921. Stat. Sci. **20**, 32–49 (2005)
1323. Stigler, S.M.: How Ronald Fisher became a mathematical statistician. Math. Soc. Sci. **176**, 23–30 (2006)
1324. Still, A.W., White, A.P.: The approximate randomization test as an alternative to the *F* test in analysis of variance. Br. J. Math. Stat. Psychol. **34**, 243–252 (1981)
1325. Stone, E.G.: On the rejection of discordant observations. Mon. Not. R. Astron. Soc. **34**, 9–15 (1873)
1326. Stuart, A.: The estimation and comparison of strengths of association in contingency tables. Biometrika **40**, 105–110 (1953)
1327. Stuart, A.: Spearman-like computation of Kendall’s tau. Br. J. Math. Stat. Psychol. **30**, 104–112 (1977)
1328. Stuart, A.: Sir Maurice Kendall, 1907–1983. J. R. Stat. Soc. A Gen. **147**, 120–122 (1984)

1329. Stuart, A., Ord, J.K., Arnold, S.: Kendall's Advanced Theory of Statistics, vol. 2A, 6th edn. Arnold, London (1999)

1330. "Student": The probable error of a correlation coefficient. Biometrika **7**, 302–310 (1908) ["Student" is a nom de plume for William Sealy Gosset]

1331. "Student": The probable error of a mean. Biometrika **6**, 1–25 (1908) ["Student" is a nom de plume for William Sealy Gosset]

1332. "Student": Untitled. Nature **124**, 93 (1929) ["Student" is a nom de plume for William Sealy Gosset]

1333. Suissa, S., Shuster, J.J.: Exact unconditional sample sizes for the 2 by 2 binomial trial. J. R. Stat. Soc. A Gen. **148**, 317–327 (1985)

1334. Suissa, S., Shuster, J.J.: The 2×2 matched-pairs trial: Exact unconditional design and analysis. Biometrics **47**, 361–372 (1991)

1335. Sun, Y.Q., Sherman, M.: Some permutation tests for survival data. Biometrics **52**, 87–97 (1996)

1336. Swade, D.: The Difference Engine: Charles Babbage and the Quest to Build the First Computer. Viking, New York (2000)

1337. Swed, F.S., Eisenhart, C.: Tables for testing randomness of grouping in a sequence of alternatives. Ann. Math. Stat. **14**, 66–87 (1943)

1338. Swofford, D.L., Thorne, J.L., Felsenstein, J., Wiegmann, B.M.: The topology-dependent permutation test for monophyly does not test for monophyly. Syst. Biol. **45**, 575–579 (1996)

1339. Taha, M.A.H.: Rank test for scale parameter for asymmetrical one-sided distributions. Publ. Inst. Stat. Paris **13**, 169–180 (1964)

1340. Taplin, S.H., Rutter, C.M., Elmore, J.G., Seger, D., White, D., Brenner, R.J.: Accuracy of screening mammography using single versus independent double interpretation. Am. J. Roentgenol. **174**, 1257–1262 (2000)

1341. Tate, M.W., Hyer, L.A.: Inaccuracy of the χ^2 test of goodness of fit when expected frequencies are small. J. Am. Stat. Assoc. **68**, 836–841 (1973)

1342. Technion: Israel Institute of Technology, Haifa, Israel: Exact Statistics. http://www.technion.ac.il/docs/sas/stat/chap28/sect28.htm. Accessed 6 Dec 2011

1343. Tedin, O.: The influence of systematic plot arrangements upon the estimate of error in field experiments. J. Agric. Sci. **21**, 191–208 (1931)

1344. Teichroew, D.: A history of distribution sampling prior to the era of the computer and its relevance to simulation. J. Am. Stat. Assoc. **60**, 27–49 (1965)

1345. ter Braak, C.J.F.: Update Notes: CANOCO Version 3.1. Agricultural Mathematics Group, Wageningen (1990)

1346. ter Braak, C.J.F.: Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel, K.H., Rothe, G., Sendler, W. (eds.) Bootstrapping and Related Techniques, pp. 79–86. Springer, Berlin (1992)

1347. Terpstra, T.J.: The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Indagat. Math. **14**, 327–333 (1952)

1348. Tezuka, S.: Uniform Random Numbers: Theory and Practice. Kluwer, Boston (1995)

1349. Thakur, A.K., Berry, K.J., Mielke, P.W.: A FORTRAN program for testing trend and homogeneity in proportions. Comput. Prog. Biomed. **19**, 229–233 (1985)

1350. The LINC, a history and restoration. DigiBarn Computer Museum. http://www.digibarn.com/stories/linc/index.html (2007). Accessed 11 Nov 2012

1351. The Royal Society: History. The Royal Society. http://royalsociety.org/about-us/history/ (2012). Accessed 28 Mar 2012

1352. The Story of the Manor of Rothamsted. Harpenden History. http://www.harpenden-history.org.uk/page_id__200_path__0p2p69p.aspx (2011). Accessed 3 Apr 2013

1353. Thisted, R.A., Velleman, P.F.: Computers and modern statistics. In: Hoaglin, D.C., Moore, D.S. (eds.) Perspectives on Contemporary Statistics, Mathematical Association of America Notes and Reports, Number 21, chap. 3, pp. 41–53. Mathematical Association of America, Washington, DC (1992)

1354. Thomas, A.J.: Dr. Winifred Brenchley, O.B.E. (1883–1953). amadon.org. http://www.amadon.org/amanda/brenchley.html (2003). Accessed 8 Nov 2012
1355. Thomas, D.G.: Algorithm 36: Exact confidence limits for the odds ratio in a 2 × 2 table. J. R. Stat. Soc. C Appl. **20**, 105–110 (1971)
1356. Thomas, D.G.: Exact and asymptotic methods for the combination of 2 × 2 tables. Comput. Biomed. Res. **8**, 423–446 (1975)
1357. Thompson, D.: Volcano Cowboys. St. Martin's Press, New York (2000)
1358. Thompson, D.W.: On Growth and Form: The Complete Revised Edition. Dover, New York (1992)
1359. Thompson, W.D., Walter, S.D.: A reappraisal of the kappa coefficient. J. Clin. Epidemiol. **41**, 949–958 (1988)
1360. Thompson, W.R.: Biological applications of normal range and associated significance tests in ignorance of original distribution forms. Ann. Math. Stat. **9**, 122–128 (1938)
1361. Thornton, H.G.: Edward John Russell: 1872–1965. Biogr. Mem. Fellows R. Soc. **12**, 456–477. http://rsbm.royalsocietypublishing.org/content/12/456.full.pdf+html (1 November 1966). Accessed 8 Nov 2012
1362. Tippett, L.H.C.: Random sampling numbers. In: Tracts for Computers, vol. 15. Cambridge University Press, Cambridge (1927)
1363. Todhunter, I.: A History of the Mathematical Theory of Probability: From the Time of Pascal to That of Laplace. Chelsea, Bronx (1965/1865) [A 1965 textually-unaltered reprint of the 1865 original]
1364. Tompkins, C.B.: Machine attacks on problems whose variables are permutations. In: Curtiss, J.H. (ed.) Numerical Analysis, vol. VI, Proceedings of Symposia in Applied Mathematics, pp. 195–211. McGraw-Hill, New York (1956)
1365. TOP500 Supercomputing Sites. http://www.top500.org (2011). Accessed 12 Mar 2012
1366. Toppo, G.: Study's rankings boost U.S. schools. USA Today. http://www.usatoday.com/news/education/story/2012-02-16/us-schools-global-ranking/53110494/1 (16 February 2012). Accessed 17 Feb 2012
1367. Tracey, D.S., Khan, K.A.: Fourth exact moment result for improving MRBP based inferences. J. Stat. Plan. Infer. **28**, 263–270 (1991)
1368. Tracey, D.S., Khan, K.A.: Fourth moment results for MRBP and related power performance. Commun. Stat. Theor. Methods **20**, 2701–2718 (1991)
1369. Tracey, T.J.G.: RANDALL: A Microsoft Fortran program for a randomization test of hypothesized order relations. Educ. Psychol. Meas. **57**, 164–168 (1997)
1370. Tritchler, D.L.: An algorithm for exact logistic regression. J. Am. Stat. Assoc. **79**, 709–711 (1984)
1371. Tritchler, D.L., Pedrini, D.T.: A computer program for Fisher's exact probability test. Educ. Psychol. Meas. **35**, 717–719 (1975)
1372. Trotter, H.F.: Algorithm 115: PERM. Commun. ACM **5**, 434 (1962)
1373. Trueman, J.W.H.: Permutation tests and outgroups. Cladistics **12**, 253–261 (1996)
1374. Tucker, D.F., Mielke, P.W., Reiter, E.R.: The verification of numerical models with multivariate randomized block permutation procedures. Meteorol. Atmos. Phys. **40**, 181–188 (1989)
1375. Tukey, J.W.: Some sampling simplified. J. Am. Stat. Assoc. **45**, 501–519 (1950)
1376. Tukey, J.W.: Bias and confidence in not-quite large samples. Ann. Math. Stat. **29**, 614 (1958)
1377. Tukey, J.W.: Approximate confidence limits for most estimates (1959) [Unpublished manuscript]
1378. Tukey, J.W.: Discussion of the papers of Messrs. Anscombe and Daniel. Technometrics **2**, 160–165 (1960)
1379. Tukey, J.W.: A survey of sampling from contaminated distributions. In: Olkin, I., Hoeffding, W., Ghurye, S.G., Madow, W.G., Mann, H.B. (eds.) Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, pp. 448–485. Stanford University Press, Stanford (1960)
1380. Tukey, J.W.: Data analysis and behavioral science (1962) [Unpublished manuscript]

1381. Tukey, J.W.: Tightening the clinical trial. Control. Clin. Trials **14**, 266–285 (1993)

1382. Tukey, J.W.: Randomization and re-randomization: The wave of the past in the future. In: Statistics in the Pharmaceutical Industry: Past, Present and Future. Philadelphia Chapter of the American Statistical Association (June 1988) [Presented at a Symposium in Honor of Joseph L. Ciminera held in June 1988 at Philadelphia, Pennsylvania]

1383. Tukey, J.W., Olmstead, P.S.: The corner test for association. Ann. Math. Stat. **18**, 299 (1947)

1384. Umesh, U.N., Peterson, R.A., Sauber, M.H.: Interjudge agreement and the maximum value of kappa. Educ. Psychol. Meas. **49**, 835–850 (1989)

1385. Upton, G.J.G.: A comparison of alternative tests for the 2 × 2 comparative trial. J. R. Stat. Soc. A Gen. **145**, 86–105 (1982)

1386. Upton, G.J.G.: Fisher's exact test. J. R. Stat. Soc. A Gen. **155**, 395–402 (1992)

1387. Ury, H.K., Kleinecke, D.C.: Tables of the distribution of Spearman's footrule. J. R. Stat. Soc. C Appl. **28**, 271–275 (1979)

1388. Valz, P.D., Thompson, M.E.: Exact inference for Kendall's S and Spearman's ρ with extension to Fisher's exact test in $r \times c$ contingency tables. J. Comput. Graph. Stat. **3**, 459–472 (1994)

1389. van den Brink, W.P., van den Brink, S.G.L.: A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. Br. J. Math. Stat. Psychol. **42**, 183–189 (1989)

1390. van der Heiden, J.A.: On a correction term in the method of paired comparisons. Biometrika **39**, 211–212 (1952)

1391. van der Reyden, D.: A simple statistical significance test. Rhod. Agric. J. **49**, 96–104 (1952)

1392. Van Helden, A.: Thomas Harriot (1560–1621). The Galileo Project. http://galileo.rice.edu/sci/harriot.html (1995). Accessed 9 Sept 2012

1393. Vanbelle, S., Albert, A.: A note on the linearly weighted kappa coefficient for ordinal scales. Stat. Methodol. **6**, 157–163 (2008)

1394. Vance, A.: Bill Joy's greatest gift to man — the vi editor. The Register. http://www.theregister.co.uk/2003/09/11/bill_joys_greatest_gift/ (11 September 2003). Accessed 6 Oct 2012

1395. Vankeerberghen, P., Vandenbosch, C., Smeyers-Verbeke, J., Massart, D.L.: Some robust statistical procedures applied to the analysis of chemical data. Chemometr. Intell. Lab. **12**, 3–13 (1991)

1396. Venkatraman, E.S.: A permutation test to compare operating characteristic curves. Biometrics **56**, 1134–1138 (2000)

1397. Verbeek, A., Kroonenberg, P.M.: A survey of algorithms for exact distributions of test statistics in $r \times c$ contingency tables with fixed margins. Comput. Stat. Data Anal. **3**, 159–185 (1985)

1398. Verdooren, L.R.: Extended tables of critical values for Wilcoxon's test statistic. Biometrika **50**, 177–186 (1963)

1399. Vollset, S.E., Hirji, K.F.: A microcomputer program for exact and asymptotic analysis of several 2 × 2 tables. Epidemiology **2**, 217–220 (1991)

1400. Vollset, S.E., Hirji, K.F., Elashoff, R.M.: Fast computation of exact confidence limits for the common odds ratio in a series of 2 × 2 tables. J. Am. Stat. Assoc. **86**, 404–409 (1991)

1401. von Eye, A., von Eye, M.: On the marginal dependency of Cohen's κ. Eur. Psychol. **13**, 305–315 (2008)

1402. von Neumann, J.: Various techniques used in connection with random digits. In: Householder, A.S., Forsyth, G.E., Germond, H.H. (eds.) The Monte Carlo Method, no. 12 in National Bureau of Standards Applied Mathematics Series, pp. 36–38. United States Government Printing Office, Washington, DC (1951)

1403. Vuong, A.: A new chip off the old block. Denver Post **120**, 1A, 16A (2 June 2013)

1404. Waksman, S.A.: The men who made Rothamsted. Soil Sci. Soc. Am. J. **8**, 5 (1944)

1405. Wald, A., Wolfowitz, J.: On a test whether two samples are from the same population. Ann. Math. Stat. **11**, 147–162 (1940)

1406. Wald, A., Wolfowitz, J.: An exact test for randomness in the non-parametric case based on serial correlation. Ann. Math. Stat. **14**, 378–388 (1943)

1407. Wald, A., Wolfowitz, J.: Statistical tests based on permutations of the observations. Ann. Math. Stat. **15**, 358–372 (1944)

1408. Walker, D.D., Loftis, J.C., Mielke, P.W.: Permutation methods for determining the significance of spatial dependence. Math. Geol. **29**, 1011–1024 (1997)

1409. Walker, H.M.: Studies in the History of Statistical Method. Williams and Wilkins, Baltimore (1929)

1410. W. Allen Wallis, obituary. Institute of Political Economy. http://www.lib.rochester.edu/index.cfm?page=4727. Accessed 20 Jan 2012

1411. Wallis, W.A.: The correlation ratio for ranked data. J. Am. Stat. Assoc. **34**, 533–538 (1939)

1412. Wallis, W.A.: The statistical research group, 1942–1945 (with comments). J. Am. Stat. Assoc. **75**, 320–330 (1980)

1413. Walters, D.E.: Sampling the randomization distribution. Statistician **30**, 289–295 (1981)

1414. Wan, Y., Cohen, J., Guerra, R.: A permutation test for the robust sib-pair method. Ann. Hum. Genet. **61**, 79–87 (1997)

1415. Warington, R.: Sir John Bennet Lawes, Bart. 1814–1900. Proc. R. Soc. Lond. **75**, 228–236 (1904)

1416. Warington, R.: Sir Joseph Henry Gilbert. 1817–1901. Proc. R. Soc. Lond. **75**, 236–242 (1904)

1417. Warnock, T.: Random-number generators. Los Alamos Sci. **15**, 137–141 (1987)

1418. Wasserstein, R.: George Box: A model statistician. Significance **7**, 134–135 (2010)

1419. Watnik, M.: Early computational statistics. J. Comput. Graph. Stat. **20**, 811–817 (2011)

1420. Weber, B., Mahapatra, S., Ryu, H., Fuhrer, A., Reusch, C.G., Thompson, D.L., Lee, W.C.T., Klimeck, G., Hollenberg, L.C.L., Simmons, M.Y.: Ohm's law survives to the atomic scale. Science **335**, 64–67 (6 January 2012)

1421. Weerahandi, S.: Exact Statistical Methods for Data Analysis. Springer, New York (1995)

1422. Wegman, E.J., Solka, J.L.: Statistical software for today and tomorrow. In: Encyclopedia of Statistics. Wiley. http://binf.gmu.edu/~jsolka/PAPERS/ess2542_rev1.pdf (2005). Accessed 13 July 2012

1423. Wei, L.J., Lachin, J.M.: Properties of the urn randomization in clinical trials. Control. Clin. Trials **9**, 345–364 (1988)

1424. Weik, M.H.: The ENIAC story. Ordnance **45**, 571–575 (1961)

1425. Weinberg, J.M., Lagakos, S.W.: Efficiency comparisons of rank and permutation tests based on summary statistics computed from repeated measures data. Stat. Med. **20**, 705–731 (2001)

1426. Weiss, L.: Wald, Abraham. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 164–167. Wiley, New York (1997)

1427. Welch, B.L.: The specification of rules for rejecting too variable a product, with particular reference to an electric lamp problem. Suppl. J. R. Stat. Soc. **3**, 29–48 (1936)

1428. Welch, B.L.: On the z-test in randomized blocks and Latin squares. Biometrika **29**, 21–52 (1937)

1429. Welch, B.L.: On tests for homogeneity. Biometrika **30**, 149–158 (1938)

1430. Welch, B.L.: The significance of the difference between two means when the population variances are unequal. Biometrika **29**, 350–362 (1938)

1431. Welch, W.J.: Rerandomizing the median in matched-pairs designs. Biometrika **74**, 609–614 (1987)

1432. Welch, W.J.: Construction of permutation tests. J. Am. Stat. Assoc. **85**, 693–698 (1990)

1433. Welch, W.J., Gutierrez, L.G.: Robust permutation tests for matched-pairs designs. J. Am. Stat. Assoc. **83**, 450–455 (1988)

1434. Wellington, J.F., Narula, S.C.: Variable selection in multiple linear regression using the minimum sum of weighted absolute errors criterion. Commun. Stat. Simul. C **B10**, 641–648 (1981)

1435. Wells, M.B.: Generation of permutations by transposition. Math. Comput. **15**, 192–195 (1961)
1436. Wells, M.B.: Computing at LASL in the 1940s and 1950s: MANIAC. Tech. rep., Los Alamos Scientific Laboratory, Los Alamos (May 1978)
1437. Westfall, P.H., Young, S.S.: Resampling-based Multiple Testing: Examples and Methods for *p*-value Adjustment. Wiley, New York (1993)
1438. Westgard, J.O., Hunt, M.R.: Use and interpretation of common statistical tests in method-comparison studies. Clin. Chem. **19**, 49–56 (1973)
1439. Westlund, K.B., Kurland, L.T.: Studies on multiple sclerosis in Winnipeg, Manitoba and New Orleans, Louisiana. Am. J. Hyg. **57**, 380–396 (1953)
1440. Wheldon, M.C., Anderson, M.J., Johnson, B.W.: Identifying treatment effects in multichannel measurements in electroencephalographic studies: Multivariate permutation tests and multiple comparisons. Aust. N. Z. J. Stat. **49**, 397–413 (2007)
1441. White, C.: The use of ranks in a test of significance for comparing two treatments. Biometrics **8**, 33–41 (1952)
1442. White, C.: The committee problem. Am. Stat. **25**, 25–26 (1971)
1443. Whitfield, J.W.: Rank correlation between two variables, one of which is ranked, the other dichotomous. Biometrika **34**, 292–296 (1947)
1444. Whitfield, J.W.: Uses of the ranking method in psychology. J. R. Stat. Soc. B Met. **12**, 163–170 (1950)
1445. Whitney, D.R.: A Comparison of the Power of Non-parametric Tests and Tests Based on the Normal Distribution Under Nonnormal Alternatives (1948) [Unpublished Ph.D. dissertation at The Ohio State University, Columbus, Ohio]
1446. Whitworth, W.A.: Choice and Chance. G. E. Stechert, New York (1942)
1447. Who was Charles Babbage? Babbage Institute, University of Minnesota. http://www.cbi.umn.edu/about/babbage.html (2011). Accessed 28 Mar 2012
1448. Wilcox, R.R.: Some results on the Tukey–McLaughlin and Yuen methods for trimmed means when distributions are skewed. Biometrical J. **36**, 259–273 (1993)
1449. Wilcox, R.R.: A one-way random-effects model for trimmed means. Psychometrika, 289–306 (1994)
1450. Wilcox, R.R.: Introduction to Robust Estimation and Hypothesis Testing. Academic, San Diego (1997)
1451. Wilcox, R.R.: Applying Contemporary Statistical Techniques. Academic, San Diego (2003)
1452. Wilcox, R.R., Keseleman, H.J., Muska, J., Cribbie, R.: Repeated measures ANOVA: Some new results on comparing trimmed means and means. Br. J. Math. Stat. Psychol. **53**, 69–82 (2000)
1453. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bull. **1**, 80–83 (1945)
1454. Wilcoxon, F.: Probability tables for individual comparisons by ranking methods. Biometrics **3**, 119–122 (1947)
1455. Wilkes, M.V.: Memoirs of a Computer Pioneer. MIT Press, Cambridge (1985)
1456. Wilks, S.S.: Order statistics. Bull. Am. Math. Soc. **54**, 6–50 (1948)
1457. Williams, E.J.: Pitman, Edwin James George. In: Johnson, N.L., Kotz, S. (eds.) Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present, Wiley Series in Probability and Statistics, pp. 153–155. Wiley, New York (1997)
1458. Williams, E.J.: Edwin James George Pitman. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 468–471. Springer, New York (2001)
1459. Williams, R.H., Zimmerman, D.W., Zumbo, B.D., Ross, D.: Charles Spearman: British behavioral scientist. Hum. Nature Rev. **3**, 114–118 (2003)
1460. Willke, T.: In Memoriam – Ransom Whitney. The Ohio State University, Department of Statistics News **16**, 8. http://www.stat.osu.edu/sites/default/files/news/statnews2008.pdf (2008). Accessed 17 Jan 2012
1461. Wilson, E.B.: The controlled experiment and the four-fold table. Science **93**, 557–560 (1941)
1462. Wilson, H.G.: Least squares versus minimum absolute deviations estimation in linear models. Dec. Sci. **9**, 322–325 (1978)

1463. Wines, M.: Foul algae follow rains for an ailing Lake Erie. NY Times **162**, A1, A14 (2013)
1464. Wolfowitz, J.: Additive partition functions and a class of statistical hypotheses. Ann. Math. Stat. **13**, 247–279 (1942)
1465. Wolfowitz, J.: On the theory of runs with some applications to quality control. Ann. Math. Stat. **14**, 280–288 (1943)
1466. Wolfowitz, J.: Non-parametric statistical inference. In: Neyman, J. (ed.) Proceedings of the [First] Berkeley Symposium on Mathematical Statistics and Probability, pp. 93–113. University of California Press, Berkeley (1949)
1467. Wolfson, M.L., Wright, H.V.: Algorithm 160: Combinatorial of M things taken N at a time. Commun. ACM **6**, 161 (1963)
1468. Wong, R.K.W., Chidambaram, N., Mielke, P.W.: Application of multi-response permutation procedures and median regression for covariate analyses of possible weather modification effects on hail responses. Atmos. Ocean **21**, 1–13 (1983)
1469. Woodall, A.D.: Generation of permutation sequences. Comput. J. **20**, 346–349 (1977)
1470. Wright, H.V., Wolfson, M.L.: Algorithm 161: Combinatorial of M things taken one at a time, two at a time, up to N at a time. Commun. ACM **6**, 161 (1963)
1471. Yamada, T., Sugiyama, T.: On the permutation test in canonical correlation analysis. Comput. Stat. Data Anal. **50**, 2111–2123 (2006)
1472. Yates, F.: Contingency tables involving small numbers and the χ^2 test. Suppl. J. R. Stat. Soc. **1**, 217–235 (1934)
1473. Yates, F.: Discussion of "Statistical problems in agricultural experimentation" by J. Neyman. Suppl. J. R. Stat. Soc. **2**, 161–166 (1935)
1474. Yates, F.: Sir Ronald Fisher and the design of experiments. Biometrics **20**, 307–321 (1964)
1475. Yates, F.: A fresh look at the basic principles of the design and analysis of experiments. In: Le Cam, L.M., Neyman, J. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. IV, pp. 777–790. University of California Press, Berkeley (1967)
1476. Yates, F.: Tests of significance for 2 × 2 contingency tables (with discussion). J. R. Stat. Soc. A Gen. **147**, 426–463 (1984)
1477. Yates, F., Mather, K.: Ronald Aylmer Fisher: 1890–1962. Biogr. Mem. Fellows R. Soc. **9**, 91–129 (1963) [Until 1955, Biographical Memoirs of Fellows of the Royal Society was known as Obituary Notices of Fellows of the Royal Society]
1478. Youden, W.J.: Randomization and experimentation. Technometrics **14**, 13–22 (1972)
1479. Yu, J., Kepner, J.L., Iyer, R.: Exact tests using two correlated binomial variables in contemporary cancer clinical trials. Biometrical J. **51**, 899–914 (2009)
1480. Yule, G.U.: On the association of attributes in statistics: With illustrations from the material childhood society. Philos. Trans. R. Soc. Lond. **194**, 257–319 (1900)
1481. Yule, G.U.: On the methods of measuring association between two attributes. J. R. Stat. Soc. **75**, 579–652 (1912)
1482. Yule, G.U., Kendall, M.G.: An Introduction to the Theory of Statistics. Griffin, London (1937)
1483. Zabel, S.L.: Ronald Aylmer Fisher. In: Heyde, C.C., Seneta, E. (eds.) Statisticians of the Centuries, pp. 389–397. Springer, New York (2001)
1484. Zabell, S.: A conversation with William Kruskal. Stat. Sci. **9**, 285–303 (1994)
1485. Zabell, S.: Meyer Dwass 1923–1996. B. Inst. Math. Stat. **27**, 1–67 (1998)
1486. Zar, J.H.: A fast and efficient algorithm for the Fisher exact test. Behav. Res. Methods Instrum. C **19**, 413–414 (1987)
1487. Zelen, M.: The analysis of several 2 × 2 contingency tables. Biometrika **58**, 129–137 (1971)
1488. Zelterman, D.: Goodness-of-fit tests for large sparse multinomial distributions. J. Am. Stat. Assoc. **82**, 624–629 (1987)
1489. Zelterman, D., Chan, I.S., Mielke, P.W.: Exact tests of significance in higher dimensional tables. Am. Stat. **49**, 357–361 (1995)
1490. Zerkowski, J.A., Powers, E.T., Kemp, D.S.: A permutation test for stabilization of polypeptide helices by sequence-dependent side chain interactions: Characterization of a helix

initiation side within the myohemerythrin sequence 76–87. J. Am. Chem. Soc. **119**, 1153–1154 (1997)

1491. Zhang, S.: The split sample permutation t-tests. J. Stat. Plan. Infer. **139**, 3512–3524 (2009)
1492. Zhou, C.: A permutation-generating algorithm. Comput. Math. Appl. **20**, 39–42 (1990)
1493. Zieffler, A.S., Harring, J.R., Long, J.D.: Comparing Groups: Randomization and Bootstrap Methods Using R. Wiley, Hoboken (2011)
1494. Zimmerman, G.M., Goetz, H., Mielke, P.W.: Use of an improved statistical method for group comparisons to study effects of prairie fire. Ecology **66**, 606–611 (1985)
1495. Zimmermann, H.: Exact calculation of permutational distributions for two dependent samples I. Biometrical J. **3**, 349–352 (1985)
1496. Zimmermann, H.: Exact calculation of permutational distributions for two independent samples. Biometrical J. **4**, 431–434 (1985)
1497. Zusne, L.: Names in the History of Psychology: A Biographical Sourcebook. Wiley, New York (1975)
1498. Zwick, R.: Another look at interrater agreement. Psychol. Bull. **103**, 374–378 (1988)

Name Index

An “n” following a page number indicates an entry contained within a footnote on that page, a **bold** number indicates a page with important information about the entry, an *italic* number indicates an entry in a figure or table heading, while a page number in Roman type indicates a textural reference.

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9,
© Springer International Publishing Switzerland 2014

Subject Index

K.J. Berry et al., *A Chronicle of Permutation Statistical Methods*,
DOI 10.1007/978-3-319-02744-9,

© Springer International Publishing Switzerland 2014