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,

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

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

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.,

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

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

$$\binom{6}{4}\binom{6}{4} p^8 (1-p)^{12-8} = 225p^8(1-p)^4 ,$$

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

and

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

thus yielding a total of

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

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

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

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

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 Bayesian 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

$$P\{1|6,6,12\} + P\{0|6,6,12\}$$
$$= \frac{6!\,6!\,6!\,6!}{12!\,1!\,5!\,5!\,1!} + \frac{6!\,6!\,6!\,6!}{12!\,0!\,6!\,6!\,0!} = \frac{36}{924} + \frac{1}{924} = \frac{37}{924}\,.$$

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

Pot	Crossed-fertilized	Self-fertilized	Difference (inches)	Difference (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

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

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

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

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 postgraduate 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

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

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]

Arrays of Σ values for $n = 1, \ldots, 5$										
1										
	1									
1	1									
	1	1								
		1	1							
1	2	2	1							
	1	2	2	1						
		1	2	2	1					
			1	2	2	1				
1	3	5	6	5	3	1				
	1	3	5	6	5	3	1			
		1	3	5	6	5	3	1		
			1	3	5	6	5	3	1	
				1	3	5	6	5	3	1
1	4	9	15	20	22	20	15	9	4	1

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.