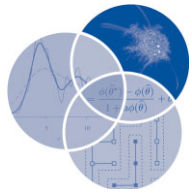
Overview



Permutation methods

Kenneth J. Berry,[1]* Janis E. Johnston[2†] and Paul W. Mielke, Jr[3]

Permutation tests are a paradox of old and new. Permutation tests pre-date most traditional parametric statistics, but only recently have become part of the mainstream discussion regarding statistical testing. Permutation tests follow a permutation or 'conditional on errors' model whereby a test statistic is computed on the observed data, then (1) the data are permuted over all possible arrangements of the data—an exact permutation test; (2) the data are used to calculate the exact moments of the permutation distribution—a moment approximation permutation test; or (3) the data are permuted over a subset of all possible arrangements of the data—a resampling approximation permutation test. The earliest permutation tests date from the 1920s, but it was not until the advent of modern day computing that permutation tests became a practical alternative to parametric statistical tests. In recent years, permutation analogs of existing statistical tests have been developed. These permutation tests provide noteworthy advantages over their parametric counterparts for small samples and populations, or when distributional assumptions cannot be met. Unique permutation tests have also been developed that allow for the use of Euclidean distance rather than the squared Euclidean distance that is typically employed in parametric tests. This overview provides a chronology of the development of permutation tests accompanied by a discussion of the advances in computing that made permutation tests feasible. Attention is paid to the important differences between 'population models' and 'permutation models', and between tests based on Euclidean and squared Euclidean distances.

 WIREs Comp Stat 2011 3 527–542 DOI: 10.1002/wics.177

Keywords: Euclidean distance; exact tests; permutation tests; resampling

INTRODUCTION

As noted by Bakeman et al.,[1] Kempthorne,[2] and Read and Cressie,[3] permutation tests are currently the gold standard against which conventional parametric tests are tested and evaluated. In this overview, permutation statistical methods are introduced and compared with conventional statistical methods, the different types of permutation methods are distinguished and described, the advantages of permutation methods are detailed, a historical chronology of the development of permutation methods is provided, and the major contributions to permutation methods are surveyed.

Essentially, two models of statistical inference exist: the population model and the permutation model.[4,5] The population model was formally proposed by Neyman and Pearson in 1928.[6,7] The population model assumes random sampling from one or more specified populations. Under the population model, the level of statistical significance that results from applying a statistical test to the results of an experiment or a survey corresponds to the frequency with which the null hypothesis would be rejected in repeated random samplings from the same specified population(s). Because repeated sampling of the true population is usually impractical, it is assumed that the sampling distribution of the test statistics under repeated random sampling conforms to an assumed theoretical distribution, such as the normal distribution. The size of a test, e.g., .01, is the probability under a specified null hypothesis that

†The views expressed in this article are those of the author and do not necessarily reflect the position or policy of the United States Department of Agriculture or the United States government.

*Correspondence to: berry@mail.colostate.edu

[1]Department of Sociology, Colorado State University, Fort Collins, CO, USA

[2]United States Department of Agriculture, Food and Nutrition Service, Alexandria, VA, USA

[3]Department of Statistics, Colorado State University, Fort Collins, CO, USA

DOI: 10.1002/wics.177

repeated outcomes based on random samples of the same size are equal to or more extreme than the observed outcome. In the population model, assignment of treatments to subjects is viewed as fixed with the stochastic element taking the form of an error that would vary if the experiment was repeated.[4] Probabilities are then calculated based on the potential outcomes of conceptual repeated draws of these errors. The model is sometimes referred to as the 'conditional on assignment' model, as the distribution used for structuring the test is conditional on the treatment assignment of the observed sample.[4]

Permutation tests were introduced by Fisher in 1926[8] and further developed by Geary in 1927,[9] Eden and Yates in 1933,[10] and Pitman in 1937[11,12] and 1938.[13] Under the permutation model, a test statistic is computed for the observed data, then the data are permuted over all possible arrangements of the observed data and the test statistic is computed for each equally likely arrangement. For clarification, an ordered sequence of n exchangeable objects $(\omega_1, \ldots, \omega_n)$ yields $n!$ equally likely arrangements of the n objects. The proportion of arrangements with test statistic values equal to or more extreme than the observed case yields the probability of the observed test statistic. In contrast to the population model, the assignment of errors to subjects is viewed as fixed, with the stochastic element taking the form of the assignment of treatments to subjects for each arrangement.[4] Probabilities are then calculated according to all outcomes associated with assignments of treatments to subjects for each case. This model is sometimes referred to as the 'conditional on errors' model, as the distribution used for structuring the test is conditional on the individual errors drawn for the observed sample.[4]

PERMUTATION TESTS

Three types of permutation tests are common: 'exact permutation tests', 'moment approximation permutation tests', and 'resampling approximation permutation tests'. Exact permutation tests enumerate all equally likely arrangements of the observed data. For each arrangement, the desired test statistic is calculated. The obtained data yield the observed value of the test statistic. The probability of obtaining the observed value of the test statistic, or a more extreme value, is the proportion of the enumerated test statistics with values equal to or more extreme than the value of the observed test statistic. As sample sizes increase, the number of possible arrangements can become very large and exact methods become impractical. For example, permuting two small samples of sizes $n_1 = n_2 = 20$ yields 137,846,528,820 arrangements of the observed data.

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

Resampling approximation permutation tests generate and examine a Monte Carlo random subset of all possible equally likely arrangements of the observed data. In the case of a resampling permutation test, the probability of obtaining the observed value of the test statistic, or a more extreme value, is the proportion of the resampled test statistics with values equal to or more extreme than the value of the observed test statistic.[14,15]

Permutation tests differ from traditional parametric tests based on an assumed population model in several ways. First, permutation tests are data dependent, in that all the information required for analysis is contained within the observed data set. Second, permutation tests do not assume any underlying theoretical distribution.[16] Third, permutation tests do not depend on the assumptions associated with traditional parametric tests, such as normality and homogeneity.[4,17] Fourth, permutation tests provide probability values based on the discrete permutation distribution of equally likely test statistic values, rather than an approximate probability value based on a theoretical distribution, such as a normal or gamma distribution. Fifth, whereas permutation tests are suitable when a random sample is obtained from a designated population, permutation tests are also appropriate for nonrandom samples, such as are common in biomedical research, as noted by Gabriel and Hall,[16] Edgington and Onghena,[18] Bear,[19] Frick,[20] and Ludbrook and Dudley.[21] Sixth, permutation tests are appropriate when analyzing entire populations, as permutation tests are not predicated on repeated random sampling from a specified population, as noted by Edgington and Onghena,[18] Ludbrook and Dudley,[21] and Holford.[22] Seventh, a permutation analog can be defined for any conventional test

statistic; thus, researchers can choose from a wide variety of permutation tests.[23] Eighth, permutation tests are ideal for very small data sets, when theoretical distribution functions may provide very poor fits. Ninth, appropriate permutation tests are resistant to extreme values, such as are common in demographic data, e.g., income, age at first marriage, number of children, etc. Consequently, any data transformation is questionable in a permutation context, e.g., square root, logarithmic, ranks, and the use of rank order statistics and the choice of a distance function, in particular, may be very misleading.[24] Finally, permutation tests provide data-dependent statistical inferences only to the actual experiment or survey that has been performed and are not dependent on a contrived super population.

This overview is confined to permutation methods, although many researchers consider that permutation methods and bootstrapping are closely related. While permutation methods and bootstrapping both involve computing simulations, and the rejection of the null hypothesis occurs when a common test statistic is extreme under both bootstrapping and permutation, they are conceptually and mechanically quite different. The two approaches differ in their distinct sampling methods. In resampling, a 'new' sample is obtained by drawing the data without replacement, whereas in bootstrapping a 'new' sample is obtained by drawing from the data with replacement.[4,25] In addition, when bootstrapping is used with small samples it is necessary to make complex adjustments to control the risk of error; see, e.g., Efron and Tibshirani,[26] Hall and Wilson,[27] and Westfall and Young.[28] Thus, bootstrapping remains firmly in the conditional-on-assignment tradition assuming that the true error distribution can be approximated by a discrete distribution with equal probability attached to each of the cases. On the other hand, permutation tests view the errors as fixed in repeated samples.[4] Finally, some researchers have tacitly conceived of permutation methods in a Bayesian context. Specifically, this interpretation amounts to a primitive Bayesian analysis where the prior distribution is the assumption of equally likely arrangements associated with the observed data, and the posterior distribution is the resulting data-dependent distribution of the test statistic induced by the prior distribution.

A CHRONOLOGY OF PERMUTATION METHODS

1920–1939: The Beginnings of Permutation Methods

The earliest discussions of permutation tests appeared in the literature in the 1920s. Splawa-Neyman[29] foreshadowed the use of permutation tests in a 1923 article, although there is no indication that any of those who worked to establish the field of permutation methods were aware of the work by Neyman, which was not translated from its original Polish language text until 1990. In this 1923 article, Neyman introduced a model for the analysis of field experiments conducted for the purpose of comparing a number of crop varieties.[29] Two years later, in *Statistical Methods for Research Workers* (Ref 8, Chapter 5, Section 24, Example 19), Fisher calculated an exact probability using the binomial probability distribution. Although using the binomial distribution to obtain a probability value is not a permutation test *per se*, the binomial distribution does yield an exact probability value. In 1927, Geary[9] first used 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.

Like Geary,[9] Eden and Yates[10] utilized permutation methods to compare a theoretical distribution to an empirical distribution. In 1933, Eden and Yates examined height measurements of wheat grown in eight blocks, each consisting of four sub-blocks of eight plots. 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 four treatment groups and four treatment blocks for a total of $(4!)^7 = 4{,}586{,}471{,}424$ possible arrangements. Eden and Yates chose a sample of 1000 of these arrangements at random and generated a table listing the simulated probabilities generated by the random sample and the theoretical counterparts to those probability values based on the normality assumption. The simulated and theoretical probabilities 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.[10]

In 1934, Fisher presented a paper describing the logic of a permutation test to the Royal Statistical Society that appeared in the *Journal of the Royal Statistical Society* the following year.[30] Fisher did not expressly discuss permutation tests, but instead used the binomial distribution to arrive at an exact probability for a 2×2 contingency table. The point of the example—that for small samples exact tests are possible, thereby eliminating the need for estimation—is indicative of an early understanding of the superiority of exact probability values computed on known discrete distributions over approximations based on theoretical distributions.

In *The Design of Experiments* in 1935 (Ref 31, Chapter 2, Section 11), Fisher again intimated at the usefulness of a permutation approach to obtain exact probabilities, and it is this text that many researchers refer to as setting the idea of permutation tests into motion, e.g., Conover,[32] Kempthorne,[33] Kruskal and Wallis,[34] and Wald and Wolfowitz.[35] Fisher's description of the 'lady tasting tea' is often used to describe the underlying logic of permutation tests. In what Fisher termed a hypothetical experiment in *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.[31] He concocted an experiment whereby the 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. 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 (Refs 31, pp. 11–29; 36, pp. 134–135; 37, pp. 1–2).

Fisher provided a second hypothetical discussion of permutation tests in *Design of Experiments* (Ref 31, Section 21), describing a way to compare the means of randomized pairs of observations by permutation. In this case Fisher carried the example through, calculating test statistics for all possible pairs of the data. For this example, Fisher considered data from Charles Darwin on 15 pairs of planters containing *Zea mays* seeds in similar soils and locations, with heights to be measured when the plants reached a given age.[38] Fisher calculated the exact probability values 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 plants. The exact probability value was calculated as the proportion of values whose differences were as, or more extreme than, the observed value. 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 (Ref 31, pp. 45–46).

Fisher's 1936 article '"The coefficient of racial likeness" and the future of craniometry' provided an alternative explanation of how permutation tests work.[39] Without calling the technique a permutation test, Fisher described a shuffling procedure for analyzing data. His description began with two groups of $n = 100$ members each and a measurement of interest on each member of the two groups. 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. 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?' (Ref 39, p. 486).

In 1936, Hotelling and Pabst[40] used permutation methods to calculate exact probabilities for small samples of ranked data in their discussion of correlation. This important article utilized the calculation of a probability that incorporated all permutations of the data, under the null hypothesis that all permutations were equally likely. The probability for any particular value was calculated as a proportion of the number of permutations equal to, or more extreme than, the value obtained from the observed data. 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, e.g., $n!$, and specific examples for small data sets. Thus, this 1936 article may well be the first example that specifically detailed the method of calculating a permutation test using all possible arrangements of the data.

The work by Hotelling and Pabst[40] became important in the discussion of distribution-free procedures involving ranked data. Fisher, however, continued to be influential in the discussion of permutation methods. Welch[41] 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. Pearson[42] also referenced the Fisher text in his consideration of randomizations with 'the lady tasting tea', but as with Fisher, neither Welch nor Pearson fully explained the technique. It was not until 1937 and 1938 that a series of articles by Pitman[11–13] explicitly discussed the permutation approach for statistical analysis; these three articles extended permutation methods to include data that were not amenable to ranking.

In the introduction to a 1937 paper on 'Significance tests which may be applied to samples from any populations', Pitman first stated that the objective 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 explicit in all of Fisher's writings' (Ref 11, p. 119). Edgington, however, noted 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, Ref 43, p. 18). Pitman further developed the permutation approach for the correlation coefficient 'which makes no assumptions about the population sampled' in the second of the three papers (Ref 12, p. 232), and then proposed a permutation test for the analysis of variance 'which involves no assumptions of normality' in the third paper (Ref 13, p. 335).

In a 1938 article, 'On tests for homogeneity,' Welch advocated calculating exact values on a limited population before moving into an examination of the moments of an infinite population.[44] Welch continued with an example of an exact calculation and further concluded that if the variances of different samples were markedly different, normal theory could badly underestimate significant differences that might exist. An exact test, however, being free from the assumptions usually associated with asymptotic statistical tests, had no such limitation.

McCarthy[45] 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[31] and 1936[39] as well as by Welch in 1938.[44] Kendall incorporated exact probabilities utilizing the 'entire universe' of permutations in the construction of τ, a new measure of rank correlation.[46] In 1939, Kendall et al.[47] utilized permutations in their discussion of Spearman's rank order correlation coefficient, and in 1938 and 1939 Olds[48] and Kendall et al.[47] calculated exact probabilities up to $n = 10$ for Spearman's rank order correlation coefficient.[49,50] The probabilities were based on their relative frequencies in the $n!$ permutations of one ranking against the other. This ushered in the 1940s when tables were published for a number of statistics with small sample sizes. These latter works constituted a harbinger of much of the work on permutations during the 1940s: a focus on creating tables for small samples that employed permutations for the calculations of exact probabilities, primarily for rank tests.

1940–1959: Interregnum

The 1940s and 1950s saw a proliferation of nonparametric rank tests, including the Kendall rank order correlation coefficient,[46,51] the Friedman two-way analysis of variance by ranks,[52,53] which is equivalent to the Kendall coefficient of concordance[51,54] (Ref 55, p. 335), the Wilcoxon rank sum test,[56] independently developed by Mann and Whitney[57] and Festinger,[58] the Wald–Wolfowitz runs test,[59] the Jonckheere–Terpstra test for ordered alternatives,[60,61] the Mann test for trend,[62] the Kruskal–Wallis one-way analysis of variance by ranks,[34] and the Mood median test.[63] In addition, permutation methods were often employed to generate tables of exact probabilities for small samples, e.g., tables for testing randomness by Swed and Eisenhart,[64] exact tables for 2×2 contingency tables by Finney,[65] exact tables for the Spearman rank order correlation coefficient by David et al.,[66] exact tables for the Wilcoxon test by Wilcoxon[56,67] and Fix and Hodges,[68] exact tables for the Mann test for trend by Mann,[62] and exact tables for the Mann–Whitney statistic by White,[69] Van den Reyden,[70] and Auble.[71]

All this is not to say, however, that theoretical work did not continue during this period. Between 1920 and 1939, emphasis was focused on validating the robustness of conventional asymptotic statistics when assumptions were violated. A theme that was commonly repeated during the 1940 to 1959 period involved difficulty of computation and, in response, conversion of data to ranks to simplify computation. In this regard, Scheffé[72] introduced nonparametric randomization tests, building on the work of Fisher.[8] Scheffé (Ref 72, p. 311) noted that 'except for very small samples the calculation... is usually extremely tedious', a problem that plagued permutation tests until the advent of high-speed computers. In that same year, Wald and Wolfowitz[73] developed an exact test procedure for randomness based on serial correlation. A year later, Wald and Wolfowitz[35] devised exact tests of significance when the form of the underlying probability distribution was unknown.

Pitman[74] in unpublished, but widely circulated, lecture notes for a course given at Columbia University in 1948 showed that the Wilcoxon[56] test for location had an asymptotic efficiency of $3/\pi$ relative to Student's t under the assumption of normality. In 1949, Wolfowitz[75] surveyed a number of problems in nonparametric inference and recommended that methods of obtaining critical regions be developed in connection with the randomization method of Fisher[8] and Pitman.[11] Freeman and Halton[76] described an exact test for small samples when chi-squared is not applicable for $r \times c$ and $2 \times 2 \times 2$ contingency tables. In 1952, Hoeffding[77] investigated the power of a family of nonparametric tests based on permutations of observations, finding the permutation tests to be asymptotically as powerful as the related parametric tests. This was a recurring theme that was also addressed by Silvey in 1955, who further considered the problem of determining the conditions under which the permutation distribution of a statistic and its normal theory distribution were asymptotically equivalent.[78] In the same year, 1955, Kempthorne described the use of randomization in experimental designs and how randomization permits evaluation

of the experimental results.[33] Included were the completely randomized design, randomized blocks, and Latin squares. Two years later, in 1957, Dwass continued the general theme of computational difficulties for permutation tests, even with small samples.[14] Dwass further recommended taking a random sample of all possible permutations for a two-sample test and making the decision to accept or reject the null hypothesis on the basis of these random permutations only.

1960–1979: Computers

Permutation tests are, by their very nature, computationally intensive and it took the development of high-speed computers for permutation tests to achieve their potential. Thus, the parallel development of permutation tests and computers is an essential part of the chronology of permutation methods. In the period prior to 1960, computers were large, slow, and expensive, and in large part their use was restricted to military and industrial applications. In the 1960s, mainframe computers became widely available to researchers at major research universities. By the end of this period, 1979, personal computers, although not common, were available to many researchers. In addition, the speed of computing increased greatly between 1960 and 1979. All this paved the way for the rapid development of permutation tests.

Permutation tests ultimately depend on the efficient generation of permutation sequences. In the case of exact permutation tests, all possible permutation sequences are generated; but for resampling permutation tests, only a random sample of permutation sequences is required. Although the first explicit description of computer algorithms for the generation of permutation sequences was given by Tompkins in 1956,[79] the period 1960–1979 was when many algorithms were presented for the generation of permutation sequences, each touting increased speed or efficiency.

While computer algorithms to generate permutation sequences were important, other researchers turned their attention to computing exact probability values for known statistics. Gregory[80] and Tritchler and Pedrini,[81] for example, confined their applications to the Fisher exact test for 2×2 contingency tables, while Agresti and Wackerly,[82] Agresti et al.,[83] Fleishman,[84] Howell and Gordon,[85] and March[86] attempted to extend the Fisher hypergeometric procedure to larger $r \times c$ contingency tables, and others applied permutation procedures to, for example, the Pitman test for two independent samples,[87] the F test for completely randomized designs,[88] the F test for randomized block designs,[89] the chi-squared test for goodness-of-fit,[90] the Kruskal–Wallis test for ranks,[91] and alternative choices of rank scores.[92,93] In addition, Edgington provided permutation procedures and examples for an extensive inventory of statistical tests in 1969 (Ref 94, pp. 93–159), and 10 years later Boyett presented an important resampling algorithm for $r \times c$ contingency tables.[95]

In 1976, Mielke et al.[96] introduced multi-response permutation procedures (MRPP), the first statistics designed especially for permutation methods, in contrast to permutation alternatives to conventional tests. On the basis of Euclidean distances, rather than squared Euclidean distances, MRPP provided exceedingly robust, distribution-free, Euclidean-based permutation alternatives to experimental designs that normally employed conventional ANOVA or MANOVA analyses.[24,97]

Between 1960 and 1979, researchers were focused on defining efficient methods for computing probability values. Existing inefficiencies were largely due to inadequate numerical algorithms, low computer clock speeds, small and slow core memories, and inefficient data transfer. Mielke et al.[96] and Mielke[98] introduced moment approximation permutation procedures whereby implementation of symmetric means, first introduced by Tukey,[99] provided the exact first three moments of a continuous distribution that approximated the discrete permutation distribution. The moment approximation permutation procedure immediately eliminated many of the computing difficulties that had plagued the computation of permutation probability values, provided an approximation to the underlying permutation distribution, and circumvented the extensive calculations of an exact permutation approach.

1980–1999: Arrival

Permutation tests arrived at a level of maturity during the period between 1980 and 1999 primarily as a result of two factors: greatly improved computer clock speeds and widely available desktop computers. Interest did continue in the study of linear rank order statistics.[100] At the same time, there was a dramatic shift in this period in sources of permutation publications. In the previous period, 1960–1979, nearly all published papers on permutation methods appeared in computer journals, such as *Communications of the ACM* and *The Computer Journal*. In the period 1980–1999, there was a shift away from computer journals into statistical journals, such as the *Journal of the American Statistical Association* and *Applied*

Statistics. An increasing number of published papers on permutations began appearing in discipline journals, such as *Educational and Psychological Measurement*, *Econometrica*, *Ecology*, *Behavior Research Methods, Instruments, & Computers*, the *Journal of Applied Meteorology*, and *Vegetatio*. In addition, a number of books on permutation methods appeared in this period, beginning with the first edition of Edgington's *Randomization Tests* in 1980,[101] a second edition 7 years later in 1987,[102] and a third edition in 1995.[43] Edgington's book was quickly followed by Hubert's 1987 book on *Assignment Methods in Combinatorial Data Analysis*,[103] Good's two books in 1994 on *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*[104] and *Permutation, Parametric and Bootstrap Tests of Hypotheses*,[105] Manly's first edition of *Randomization and Monte Carlo Methods in Biology* in 1991[106] followed by a second edition in 1997,[107] Weerahandi's book in 1995 on *Exact Statistical Methods for Data Analysis*,[108] and Good's third book on *Resampling Methods: A Practical Guide to Data Analysis* in 1999.[109]

During this period, work also continued on improving the computational efficiency of permutation tests, inspired by the ease of calculations due to increases in computer speed and storage. Between 1980 and 1999 a number of 'algorithmic tricks' were developed that substantially reduced computation time.[16,110] Balmer[111] and Dallal[112] utilized recursive routines to efficiently generate both statistics and probability values, and Thakur et al.,[113] Berry and Mielke,[114–116] and Berry et al.[117,118] enhanced the procedure by coupling recursive routines with the use of an arbitrary origin. A second algorithmic innovation was to recognize that only the variable part of a statistical formula needed to be computed for each permutation. But, by far, the most important innovation was the introduction of the network algorithm by Mehta and Patel.[119]

In 1980, Mehta and Patel[119] introduced a network algorithm that proved to be an efficient method to calculate exact permutation tests. Originally designed for exact tests on $2 \times k$ contingency tables, the algorithm was quickly extended to the more general problem of $r \times c$ contingency tables by Pagano and Taylor Halvorsen[120] and Mehta and Patel.[121] Interest continued in this period on computational methods for both exact and resampling analyses of $r \times c$ contingency tables with articles by Balmer,[111] Berry and Mielke,[114,115,122–124] Pagano and Taylor Halvorsen,[120] Patefield,[125] Saunders,[126] Mielke and Berry,[127,128] and Baglivo et al.[129] Extensions to multidimensional tables were provided by Kreiner,[130] Mielke and Berry,[131–133] Berry and Mielke,[134] Mielke et al.,[135] and Zelterman et al.[136]

In the period between 1980 and 1999, permutation tests branched out from their home in statistics to include a variety of other disciplines, most notably in psychology with articles by Berry and Mielke[124,137–139] and Mielke and Berry[132,140,141]; pharmacology and physiology with an important article by Ludbrook[5]; biomedical sciences with articles by Ludbrook and Dudley,[21] Dallal,[112] and Zimmerman[142,143]; anthropology with articles by Mielke et al.[144] and Berry et al.[145]; ecology with articles by Zimmerman et al.[146] and Biondini et al.[147]; and atmospheric science with articles by Mielke et al.,[148] Mielke,[149,150] Wong et al.,[151] Tucker et al.,[152] Lee et al.,[153] Kelly et al.,[154] Cotton et al.,[155] and Mielke et al.[156,157]

While many of the contributions to the permutation literature during this period concentrated on efficient means to calculate permutation versions of existing statistics, the advancements in computational efficiency allowed for the development of a wider variety of statistical tests, tailored to the peculiarities of the problem under consideration. Consequently, a few researchers utilized the permutation structure to develop new statistical measures and tests. Permutation versions of existing statistics included Fisher's exact probability test by Verbeek and Kroonenberg,[110] Berry and Mielke,[115,137,139] Mehta and Patel,[121,158,159] Mielke and Berry,[128] Baglivo et al.,[129] Joe,[160] and Zar[161]; the chi-squared test by Mielke and Berry,[127] Baglivo et al.,[129] and Romesburg et al.[162]; various goodness-of-fit tests by Baglivo et al.,[129] Mielke and Berry,[163] and Tritchler[164]; the Kolmogorov–Smirnov test by Romesburg et al.[162]; the Wilcoxon signed-ranks test by Dallal[112] and Zimmerman[142]; the Wilcoxon–Mann–Whitney (WMW) test by Dallal[112] and Zimmerman[143]; the likelihood-ratio test by Baglivo et al.[129]; one-way analysis of variance by Berry and Mielke[137]; the odds ratio by Vollset and Hirji[165] and Vollset et al.[166]; the Goodman–Kruskal tau measure of nominal contingency by Berry and Mielke[114,167]; Cohen's kappa measure of agreement by Berry and Mielke[168]; logistic regression by Hirji et al.[169] and Tritchler[164]; various two-sample tests by Zimmerman,[142,143] Baker and Tilbury,[170] Chen and Dunlap,[171] and Edgington and Khuller[172]; the McNemar test by Baker and Tilbury[170]; Cochran's Q test by Mielke and Berry[173]; and the Cochran–Armitage test for trend by Mehta et al.[174]

At the same time, Mielke and his collaborators focused their work on designing permutation tests

that were not simply permutation versions of existing statistics. Conventional statistical tests and measures, both parametric and nonparametric, are based on squared Euclidean distances between data points. Examples include two-sample *t* tests, various *F* tests, ordinary least-squares regression, and nonparametric tests such as the WMW rank sum test, the Kruskal–Wallis one-way analysis of variance by ranks, and the Friedman two-way analysis of variance by ranks. A Euclidean distance function based on absolute distances between data points was incorporated into new permutation tests for matched pairs designs by Mielke and Berry,[140] Berry and Mielke,[175,176] Brockwell and Mielke,[177] Mielke and Berry,[178] and Mielke et al.[179]; completely randomized designs by Mielke et al.,[148] Berry et al.,[180,181] Berry and Mielke,[182–184] O'Reilly and Mielke,[185] Brockwell et al.,[186] Mielke,[187,188] Mielke et al.,[189] and Mielke and Berry[190]; randomized block designs by Mielke,[149] Tucker et al.,[152] Brockwell and Mielke,[177] Mielke and Berry,[178] Berry and Mielke,[191] and Mielke and Iyer[192]; contingency table analyses by Berry and Mielke,[114,122,134,167,193,194] Mielke,[195] Mielke and Berry,[131] and Zelterman et al.[136]; goodness-of-fit tests by Mielke and Berry[163] and Berry and Mielke[196]; multiple regression by Mielke and Berry[197,198] and Berry and Mielke;[184,191,199–201] and measures of agreement and consensus by Berry and Mielke.[168,193,194,202,203]

2000–2010: Maturity

Clock speeds on personal computers increased significantly between 2000 and 2010. In 2000, the Intel Pentium processor contained 42 million transistors and ran at 1.5 GHz. In the spring of 2010, Intel released the Itanium processor, code named Tukwila after a town in Washington, containing 2 billion transistors and running at 4.8 GHz. While not widely available to researchers, by 2010 mainframe computers were measuring computing speeds in teraflops. To emphasize the progress of computing, in 1951 the Remington Rand Corporation introduced the Univac computer running at 1905 flops, which with ten mercury delay line memory tanks could store 20,000 bytes of information; in 2008 the IBM Corporation supercomputer, codenamed Roadrunner, reached a sustained performance of 1 petaflop; in 2010 the Cray Jaguar was named the world's fastest computer performing at a sustained speed of 1.75 petaflops with 360 terabytes of memory; and in November 2010 China exceeded the computing speed of the Cray Jaguar by 57% with the introduction of the Tianhe-A1 super computer performing at 2.67 petaflops. From a more general perspective, in 1977 the Tandy Corporation released the TRS-80, the first fully assembled personal computer, distributed through Radio Shack stores. The TRS-80 had 4 MB of RAM and ran at 1.78 MHz. By way of comparison, in 2010 the Apple iPhone had 131,072 times the memory of the TRS-80 and was about 2000 times faster, running at 1 GHz. By 2010, computing power was finally sufficient to accommodate the needs of computational statisticians running permutation tests. Keller–McNulty and Higgins[204] concluded on the basis of Monte Carlo results that there was little reason to conduct exact permutation tests, recommending that researchers use only 1600 random samples. Bailer,[205] Kim et al.,[206] and McQueen[207] used only 1000 random permutations in their studies, and Edgington[94] showed that 999 random permutations of the data were sufficient. Dwass[14] argued that 10,000 random permutations provided results nearly as powerful as complete enumeration, and Edgington and Khuller[172] concurred. Manly (Ref 106, pp. 32–36) and Noreen (Ref 208, p. 15) argued that for testing at the 5% level of significance, 1000 random permutations were sufficient. Because of increasing computing power, by 2010 probability values based on exact enumeration sometimes exceeded 10,000,000 permutations and resampling probability values based on 1,000,000 random permutations were not only recommended,[209] but common.[23]

Increased computational efficiency paved the way for the introduction of a number of software packages for permutation tests, now widely available to computational statisticians. Among the most available and popular software packages for permutation tests are Box Sampler (Microsoft Corp., Redmond, WA), S-Plus (MathSoft, Inc., Seattle, WA), Statistica (StatSoft, Inc., Tulsa, OK), SPSS (SPSS, Inc., Chicago, IL), SAS (SAS Institute, Inc., Cary, NC), Stata (StataCorp LP, College Station, TX), Blossom Statistical Software (Fort Collins Ecological Science Center, Fort Collins, CO), Resampling Stats (Resampling Stats, Inc., Arlington, VA), Statistical Calculator (StatPac, Bloomington, MN), StatXact (Cytel Software Corp., Cambridge, MA), Systat (Systat Software, Inc., Chicago, IL), and Testimate (Institute for Data Analysis and Study Planning, Munich, Germany).

In addition to permutation software, the decade 2000–2009 saw the publication of a number of books on permutation methods, including volumes on *Data Analysis by Resampling: Concepts and Applications* by Lunneborg in 2000,[210] a second edition of *Permutation Tests: A Practical Guide*

to Resampling Methods for Testing Hypotheses by Good in 2000,[211] a second edition of *Permutation, Parametric and Bootstrap Tests of Hypotheses* by Good in 2000,[212] a second edition of *Resampling Methods: A Practical Guide to Data Analysis* by Good in 2001,[213] *Permutation Methods: A Distance Function Approach* by Mielke and Berry in 2001,[214] *Multivariate Permutation Tests: With Applications in Biostatistics* by Pesarin in 2001,[215] *Resampling Methods for Dependent Data* by Lahiri in 2003,[216] a third edition of *Permutation, Parametric and Bootstrap Tests of Hypotheses* by Good in 2005,[217] a third edition of *Resampling Methods: A Practical Guide to Data Analysis* by Good in 2006,[218] *Exact Analysis of Discrete Data* by Hirji in 2006,[219] a fourth edition of *Randomization Tests* by Edgington and Onghena in 2007,[18] a third edition of *Randomization, Bootstrap and Monte Carlo Methods in Biology* by Manly in 2007,[220] and a second edition of *Permutation Methods: A Distance Function Approach* by Mielke and Berry in 2007.[23]

The journal articles on permutation methods published between 2000 and 2010 are too numerous to be summarized in any detail. A search of The Web of Science® for 'permutation' lists 9259 journal articles and 73,960 citations for this period, with steady increases for each year. For example, in 2000 there were 1619 citations, in 2005 there were 5862 citations, and in 2010 there were 15,612 citations. The journal articles may be conveniently divided into two areas: field of research and research methods.

A cursory examination of the fields of research in which articles using permutation methods were published includes biology, genetics, statistics, computer science, bioinformatics, conservation, cognition, epidemiology, ecology, medicine, history, atmospheric science, forestry, public health, environmental research, and geology.

The research methods for which permutation tests were published include, but are not limited to, multiple regression, analysis of variance, the WMW test, the Jonckheere–Terpstra test, trend analysis, matched pairs, analysis of multivariate data, partitions, Cohen's kappa, categorical variation, ordered and unordered contingency tables, qualitative variation, Cronbach's alpha, tetrachoric correlation, ridit analysis, and robustness.

CONCLUSION

Originally developed to test and confirm the robustness of conventional tests and measures such as t tests, analyses of variance, chi-squared, and correlation/regression, permutation methods have emerged as an area of statistical analysis in their own right and are the gold standard against which conventional tests are now judged.[1–3] From their inception, permutation tests were understood to be superior to conventional tests. Recall, permutation tests are data dependent, do not depend on the assumptions associated with conventional tests, are appropriate for use with a population or a nonrandom sample, and provide exact probability values. The fact that permutation tests provide exact probability values is still extremely important in verifying conventional tests. For example, Bergmann et al.[221] investigated the efficacy of a variety of statistical packages and calculated probability values for the WMW test. Utilizing a single data set, the probability of the WMW test was calculated on 11 standard statistical packages, producing a variety of very different probability values. Bergman et al. concluded that the only accurate form of the WMW test was 'one in which the exact permutation null distribution [was] compiled for the actual data' (Ref 221, p. 72). The editor of *The American Statistician* further noted that 'it is a cause of considerable concern when the results for a relatively simple test differ across packages' (Ref 222, p. 71).

The Fisher exact probability test is the iconic permutation test and is familiar to most researchers. A search of the web in early 2011 for 'Fisher exact test' yielded 1,390,000 hits. Thus, the Fisher exact probability test provides a recognizable vehicle to summarize the attributes that distinguish permutation tests from conventional tests in general.

Consider a 2×2 contingency table of n cases, where x denotes the frequency of any cell and r and c represent the row and column marginal frequency totals, respectively, corresponding to x. Given fixed marginal frequency totals, the point probability of x is equivalent to the point probability of the observed table and Fisher's exact probability is the hypergeometric point probability of x given by

$$p(x \mid n, r, c) = \frac{\binom{r}{x}\binom{n-r}{c-x}}{\binom{n}{c}} = \frac{r!\, c!\, (n-r)!\, (n-c)!}{n!\, x!\, (r-x)!\, (c-x)!\, (n-r-c+x)!}. \quad (1)$$

This, of course, is exactly the formulation of the celebrated 'lady tasting tea' experiment (Refs 31, pp. 11–29; 36, pp. 134–135; 37, pp. 1–2).

The probability of the observed table or one more extreme requires the enumerated permutation distribution of $a \le x \le b$, where $a = \max(0, r + c - n)$

and $b = \min(r, c)$. If x_o denotes the observed value of x, then the two-sided cumulative probability of x_o, with fixed marginal frequency totals, is given by

$$P(x_o \mid n, r, c) = \sum_{k=a}^{b} \Phi_k p(k \mid n, r, c), \qquad (2)$$

where

$$\Phi_k = \begin{cases} 1 & \text{if } p(k \mid n, r, c) \le p(x_o \mid n, r, c), \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

Note that this is the Fisher definition of significance probability, i.e., the probability of a given experimental observation, plus the sum of those probabilities equal to or more extreme than the observed probability (Refs 223, Chapter V, Section 7; 224, p. 486).

Obviously, the Fisher exact probability test is not included in the traditional Neyman–Pearson[6,7] population model of conditional assignment.[225,226] Indeed, the Fisher and Neyman–Pearson approaches represent two different visions of science.[224] There is no testable null hypothesis for the Fisher approach in the Neyman–Pearson sense of a posited population parameter, and no alternative hypothesis. Also, the Fisher approach contains no probability of Type I error or α; no probability of Type II error or β, and therefore no complement of the probability of Type II error, i.e., power; no point estimate of a population parameter; and, consequently, no confidence limits. For Fisher, a computed probability value was a measure of evidence in a single experiment, whereas for Neyman–Pearson, insofar as probability values are relevant, a probability value was to be interpreted as a hypothetical frequency of error if the experiment was repeated. Finally, the Fisher exact probability test is completely data dependent, makes no assumptions about a theoretical distribution, and does not require a random sample drawn from a specified population.

Early in their history, permutation methods were impractical and usually limited to the verification of conventional methods. It was the advent of high-speed computing that allowed permutation methods to become practical. Permutation methods have since supplanted conventional statistical methods for a variety of research designs, and the field continues to expand as researchers design new applications for permutation methods. Presently, it appears that computing speed is sufficient for most applications of permutation methods. When combined with resampling methods and innovative algorithms, permutation tests are a preferred alternative to many conventional statistical tests.

REFERENCES

1. Bakeman R, Robinson BF, Quera V. Testing sequential association: estimating exact *p* values using sampled permutations. *Psychol Methods* 1996, 1:4–15.
2. Kempthorne O. Some aspects of experimental inference. *J Am Stat Assoc* 1966, 61:11–34.
3. Read TRC, Cressie NAC. *Goodness-of-Fit for Discrete Multivariate Data*. New York: Springer-Verlag; 1988.
4. Kennedy PE. Randomization tests in econometrics. *J Bus Econ Stat* 1995, 13:85–94.
5. Ludbrook J. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clin Exp Pharmacol Physiol* 1994, 21:673–686.
6. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* 1928, 20A:175–240.
7. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: part II. *Biometrika* 1928, 20A:263–294.
8. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd; 1925.
9. Geary RC. Some properties of correlation and regression in a limited universe. *Metron Riv Int Stat* 1927, 7:83–119.
10. Eden T, Yates F. On the validity of Fisher's *z* test when applied to an actual example of non-normal data. *J Agric Sci* 1933, 23:6–17.
11. Pitman EJG. Significance tests which may be applied to samples from any populations. *Suppl J R Stat Soc* 1937, 4:119–130.
12. Pitman EJG. Significance tests which may be applied to samples from any populations: II. The correlation coefficient test. *Suppl J R Stat Soc* 1937, 4:225–232.
13. Pitman EJG. Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika* 1938, 29:322–335.
14. Dwass M. Modified randomization tests for nonparametric hypotheses. *Ann Math Stat* 1957, 28:181–187.
15. Hope ACA. A simplified Monte Carlo significance test procedure. *J R Stat Soc B* 1968, 30:582–598.

16. Gabriel KR, Hall WJ. Rerandomization inference on regression and shift effects: computationally feasible methods. *J Am Stat Assoc* 1983, 78:827–836.
17. Berry KJ, Mielke PW, Mielke HW. The Fisher–Pitman permutation test: an attractive alternative to the *F* test. *Psychol Rep* 2002, 90:495–502.
18. Edgington ES, Onghena P. *Randomization Tests*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC; 2007.
19. Bear G. Computationally intensive methods warrant reconsideration of pedagogy in statistics. *Behav Res Methods Instrum Comput* 1995, 27:144–147.
20. Frick RW. Interpreting statistical testing: process and propensity, not population and random sampling. *Behav Res Methods Instrum Comput* 1998, 30:527–535.
21. Ludbrook J, Dudley H. Why permutation tests are superior to *t* and *F* tests in biomedical research. *Am Stat* 1998, 52:127–132.
22. Holford TR. Editorial: exact methods for categorical data. *Stat Methods Med Res* 2003, 12:1.
23. Mielke PW, Berry KJ. *Permutation Methods: A Distance Function Approach*. 2nd ed New York: Springer-Verlag; 2007.
24. Mielke PW, Berry KJ, Johnston JE. Robustness without rank order statistics. *J Appl Stat* 2011, 38:207–214.
25. Romano JP. Bootstrap and randomization tests of some nonparametric hypotheses. *Ann Stat* 1989, 17:141–159.
26. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC; 1993.
27. Hall P, Wilson SR. Two guidelines for bootstrap hypothesis testing. *Biometrics* 1991, 47:757–762.
28. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: John Wiley & Sons; 1993.
29. Splawa-Neyman J. Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych (On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Rocz Nauk Rolniczych* 1923, 10:1–51. Translated by Dabrowska DM and Speed TP in *Stat Sci* 1990, 5:465–472.
30. Fisher RA. The logic of inductive inference. *J R Stat Soc* 1935, 98:39–82.
31. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd; 1935.
32. Conover WJ. *Practical Nonparametric Statistics*. 3rd ed. New York: John Wiley & Sons; 1999.
33. Kempthorne O. The randomization theory of experimental inference. *J Am Stat Assoc* 1955, 50:946–967.
34. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc* 1952, 47:583–621. Erratum in: *J Am Stat Assoc* 1953, 48:907–911.
35. Wald A, Wolfowitz J. Statistical tests based on permutations of the observations. *Ann Math Stat* 1944, 15:358–372.
36. Box JF. *R. A. Fisher: The Life of a Scientist*. New York: John Wiley & Sons; 1978.
37. Salsburg D. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: Henry Holt; 2001.
38. Darwin C. *The Effects of Cross and Self Fertilization in the Vegetable Kingdom*. London: John Murray; 1876.
39. Fisher RA. 'The coefficient of racial likeness' and the future of craniometry. *J R Anthropol Inst* 1936, 66:57–63.
40. Hotelling H, Pabst MR. Rank correlation and tests of significance involving no assumption of normality. *Ann Math Stat* 1936, 7:29–43.
41. Welch BL. On the *z*-test in randomized blocks and Latin squares. *Biometrika* 1937, 29:21–52.
42. Pearson ES. Some aspects of the problem of randomization. *Biometrika* 1937, 29:53–64.
43. Edgington ES. *Randomization Tests*. 3rd ed. New York: Marcel Dekker; 1995.
44. Welch BL. On tests for homogeneity. *Biometrika* 1938, 30:149–158.
45. McCarthy MD. On the application of the *z*-test to randomized blocks. *Ann Math Stat* 1939, 10:337–359.
46. Kendall MG. A new measure of rank correlation. *Biometrika* 1938, 30:81–93.
47. Kendall MG, Kendall SFH, Babington Smith B. The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika* 1939, 30:251–273.
48. Olds EG. Distribution of sums of squares of rank differences for small numbers of individuals. *Ann Math Stat* 1938, 9:133–148.
49. Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904, 15:72–101.
50. Spearman C. 'Footrule' for measuring correlation. *Br J Psychol* 1906, 2:89–108.
51. Kendall MG. *Rank Correlation Methods*. London: Griffin; 1948.
52. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 1937, 32:675–701.
53. Friedman M. A comparison of alternative tests of significance for the problem of *m* rankings. *Ann Math Stat* 1940, 11:86–92.
54. Kendall MG, Babington Smith B. The problem of *m* rankings. *Ann Math Stat* 1939, 10:275–287.
55. Savage IR. Nonparametric statistics. *J Am Stat Assoc* 1957, 52:331–344.
56. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull* 1945, 1:80–83.

57. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947, 18:50–60.
58. Festinger L. The significance of differences between means without reference to the frequency distribution function. *Psychometrika* 1946, 11:97–105.
59. Wald A, Wolfowitz J. On a test whether two samples are from the same population. *Ann Math Stat* 1940, 11:147–162.
60. Jonckheere AR. A distribution-free *k*-sample test against ordered alternatives. *Biometrika* 1954, 41:133–145.
61. Terpstra TJ. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Math* 1952, 14:327–333.
62. Mann HB. Nonparametric tests against trend. *Econometrica* 1945, 13:245–259.
63. Mood AM. On the asymptotic efficiency of certain nonparametric two-sample tests. *Ann Math Stat* 1954, 25:514–522.
64. Swed FS, Eisenhart C. Tables for testing randomness of grouping in a sequence of alternatives. *Ann Math Stat* 1943, 14:66–87.
65. Finney DJ. The FisherYates test of significance in 2×2 contingency tables. *Biometrika* 1948, 35:145–156.
66. David ST, Kendall MG, Stuart A. Some questions of distribution in the theory of rank correlation. *Biometrika* 1951, 38:131–140.
67. Wilcoxon F. Probability tables for individual comparisons by ranking methods. *Biometrics* 1947, 3:119–122.
68. Fix E, Hodges JL. Significance probabilities of the Wilcoxon test. *Ann Math Stat* 1955, 26:301–312.
69. White C. The use of ranks in a test of significance for comparing two treatments. *Biometrics* 1952, 8:33–41.
70. Van der Reyden D. A simple statistical significance test. *Rhod Agric J* 1952, 49:96–104.
71. Auble D. Extended tables for the Mann–Whitney statistic. *Bull Inst Educ Res* 1953, 1:1–39.
72. Scheffé H. Statistical inference in the non-parametric case. *Ann Math Stat* 1943, 14:305–332.
73. Wald A, Wolfowitz J. An exact test for randomness in the non-parametric case based on serial correlation. *Ann Math Stat* 1943, 14:378–388.
74. Pitman EJG. Lecture notes on nonparametric statistical inference. Unpublished notes; 1948.
75. Wolfowitz J. Non-parametric statistical inference. In: Neyman J, ed. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press; 1949, 93–113.
76. Freeman GH, Halton JH. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 1951, 38:141–149.
77. Hoeffding W. The large-sample power of tests based on permutations of observations. *Ann Math Stat* 1952, 23:169–192.
78. Silvey SD. The equivalence of asymptotic distributions under randomisation and normal theories. *Proc Glasgow Math Assoc* 1955, 1:139–147.
79. Tompkins CB. Machine attacks on problems whose variables are permutations. In: Curtiss JH, ed. *Numerical Analysis*, volume 6 of *Proceedings of Symposia in Applied Mathematics*. New York: McGraw–Hill; 1956, 195–211.
80. Gregory RJ. A FORTRAN computer program for the Fisher exact probability test. *Educ Psychol Meas* 1973, 33:697–700.
81. Tritchler DL, Pedrini DT. A computer program for Fisher's exact probability test. *Educ Psychol Meas* 1975, 35:717–719.
82. Agresti A, Wackerly D. Some exact conditional tests of independence for $R \times C$ cross-classification tables. *Psychometrika* 1977, 42:111–125.
83. Agresti A, Wackerly D, Boyett JM. Exact conditional tests for cross-classifications: approximation of attained significance levels. *Psychometrika* 1979, 44:75–83.
84. Fleishman AI. A program for calculating the exact probability along with explorations of *M* by *N* contingency tables. *Educ Psychol Meas* 1977, 37:799–803.
85. Howell DC, Gordon LR. Computing the exact probability of an *r* by *c* contingency table with fixed marginal totals. *Behav Res Methods Instrum* 1976, 8:317.
86. March DL. Exact probabilities for $R \times C$ contingency tables (Algorithm 434). *Commun ACM* 1972, 15:991–992.
87. Arbuckle J, Aiken LS. A program for Pitman's permutation test for differences in location. *Behav Res Methods Instrum* 1975, 7:381.
88. Baker FB, Collier RO. Some empirical results on variance ratios under permutation in the completely randomized design. *J Am Stat Assoc* 1966, 61:813–820.
89. Collier RO, Baker FB. Some Monte Carlo results on power of *F* tests under permutation in simple randomized block design. *Biometrika* 1966, 53:199–203.
90. Radlow R, Alf EF Jr. An alternate multinomial assessment of the accuracy of the χ^2 test of goodness of fit. *J Am Stat Assoc* 1975, 70:811–813.
91. Klotz J, Teng J. One-way layout for counts and the exact enumeration of the Kruskal–Wallis *H* distribution with ties. *J Am Stat Assoc* 1977, 72:165–169.
92. Mielke PW. Asymptotic behavior of two-sample tests based on powers of ranks for detecting scale and location alternatives. *J Am Stat Assoc* 1972, 67:850–854.
93. Mielke PW, Berry KJ. An extended class of matched pairs tests based on powers of ranks. *Psychometrika* 1976, 41:89–100.

94. Edgington ES. *Statistical Inference: The Distribution-free Approach*. New York: McGraw–Hill; 1969.
95. Boyett JM. $R \times C$ tables with given row and column totals (Algorithm 144). *Appl Stat* 1979, 28:329–332.
96. Mielke PW, Berry KJ, Johnson ES. Multi-response permutation procedures for a priori classifications. *Commun Stat Theory Methods* 1976, 5:1409–1424.
97. Mielke PW. Multiresponse permutation procedures. In: *Encyclopedia of Statistical Sciences*. Vol. V. New York: John Wiley & Sons, 1985, 724–727.
98. Mielke PW. On asymptotic non-normality of null distributions of MRPP statistics. *Commun Stat Theory Methods* 1979, 8:1541–1550. Errata in: *Commun Stat Theory Methods* 1981, 10:1795 and 1982, 11:847.
99. Tukey JW. Some sampling simplified. *J Am Stat Assoc* 1950, 45:501–519.
100. Mielke PW, Sen PK. On asymptotic non-normal null distributions for locally most powerful rank test statistics. *Commun Stat Theory Methods* 1981, 10:1079–1094.
101. Edgington ES. *Randomization Tests*. New York: Marcel Dekker; 1980.
102. Edgington ES. *Randomization Tests*. 2nd ed. New York: Marcel Dekker; 1987.
103. Hubert L. *Assignment Methods in Combinatorial Data Analysis*. New York: Marcel Dekker; 1987.
104. Good PI. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag; 1994.
105. Good PI. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer-Verlag; 1994.
106. Manly BFJ. *Randomization and Monte Carlo Methods in Biology*. London: Chapman & Hall; 1991.
107. Manly BFJ. *Randomization and Monte Carlo Methods in Biology*. 2nd ed. London: Chapman & Hall; 1997.
108. Weerahandi S. *Exact Statistical Methods for Data Analysis*. New York: Springer-Verlag; 1995.
109. Good PI. *Resampling Methods: A Practical Guide to Data Analysis*. Boston, MA: Birkhäuser; 1999.
110. Verbeek A, Kroonenberg PM. A survey of algorithms for exact distributions of test statistics in $r \times c$ contingency tables with fixed margins. *Comput Stat Data Anal* 1985, 3:159–185.
111. Balmer DW. Recursive enumeration of $r \times c$ tables for exact likelihood evaluation (Algorithm 236). *Appl Stat* 1988, 37:290–301.
112. Dallal GE. PITMAN: a FORTRAN program for exact randomization tests. *Comput Biomed Res* 1988, 21:9–15.
113. Thakur AK, Berry KJ, Mielke PW. A FORTRAN program for testing trend and homogeneity in proportions. *Comput Programs Biomed* 1985, 19:229–233.
114. Berry KJ, Mielke PW. Goodman and Kruskal's tau-b statistic: a nonasymptotic test of significance. *Sociol Methods Res* 1985, 13:543–550.
115. Berry KJ, Mielke PW. Exact chi-square and Fisher's exact probability test for 3 by 2 cross-classification tables. *Educ Psychol Meas* 1987, 47:631–636.
116. Berry KJ, Mielke PW. Exact cumulative probabilities for the multinomial distribution. *Educ Psychol Meas* 1995, 55:769–772.
117. Berry KJ, Mielke PW, Helmericks SG. Exact confidence limits for proportions. *Educ Psychol Meas* 1988, 48:713–716.
118. Berry KJ, Mielke PW, Helmericks SG. An algorithm to generate discrete probability distributions: binomial, hypergeometric, negative binomial, inverse hypergeometric, and Poisson. *Behav Res Methods Instrum Comput* 1994, 26:366–367.
119. Mehta CR, Patel NR. A network algorithm for the exact treatment of the $2 \times k$ contingency table. *Commun Stat Simul Comput* 1980, 9:649–664.
120. Pagano M, Taylor Halvorsen K. An algorithm for finding the exact significance levels of $r \times c$ contingency tables. *J Am Stat Assoc* 1981, 76:931–934.
121. Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J Am Stat Assoc* 1983, 78:427–434.
122. Berry KJ, Mielke PW. Goodman and Kruskal's tau-b statistic: a FORTRAN-77 subroutine. *Educat Psychol Meas* 1986, 46:646–649.
123. Berry KJ, Mielke PW. R by C chi-square analyses of small expected cell frequencies. *Educ Psychol Meas* 1986, 46:169–173.
124. Berry KJ, Mielke PW. Monte Carlo comparisons of the asymptotic chi-square and likelihood-ratio tests with the nonasymptotic chi-square test for sparse R by C tables. *Psychol Bull* 1988, 103:256–264.
125. Patefield WM. An efficient method of generating random $r \times c$ tables with given row and column totals (Algorithm 159). *Appl Stat* 1981, 30:91–97.
126. Saunders IA. Enumeration of $R \times C$ tables with repeated row totals (Algorithm 205). *Appl Stat* 1984, 33:340–352.
127. Mielke PW, Berry KJ. Non-asymptotic inferences based on the chi-square statistic for r by c contingency tables. *J Stat Plann Infer* 1985, 12:41–45.
128. Mielke PW, Berry KJ. Fisher's exact probability test for cross-classification tables. *Educ Psychol Meas* 1992, 52:97–101.
129. Baglivo J, Olivier D, Pagano M. Methods for the analysis of contingency tables with large and small cell counts. *J Am Stat Assoc* 1988, 83:106–113.
130. Kreiner S. Analysis of multidimensional contingency tables by exact conditional tests: techniques and strategies. *Scand J Stat* 1987, 14:97–112.

131. Mielke PW, Berry KJ. Cumulant methods for analyzing independence of *r*-way contingency tables and goodness-of-fit frequency data. *Biometrika* 1988, 75:790–793.

132. Mielke PW, Berry KJ. Nonasymptotic probability values for Cochran's *Q* statistic: a FORTRAN 77 program. *Percept Mot Skills* 1996, 82:303–306.

133. Mielke PW, Berry KJ. Exact probabilities for first-order, second-order, and third-order interactions in $2\times2\times2\times2$ contingency tables. *Percept Mot Skills* 1998, 86:760–762.

134. Berry KJ, Mielke PW. Analyzing independence in *r*-way contingency tables. *Educ Psychol Meas* 1989, 49:605–607.

135. Mielke PW, Berry KJ, Zelterman D. Fisher's exact test of mutual independence for $2\times2\times2$ cross-classification tables. *Educ Psychol Meas* 1994, 54:110–114.

136. Zelterman D, Chan IS, Mielke PW. Exact tests of significance in higher dimensional tables. *Am Stat* 1995, 49:357–361.

137. Berry KJ, Mielke PW. Moment approximations as an alternative to the *F* test in analysis of variance. *Br J Math Stat Psychol* 1983, 36:202–206.

138. Berry KJ, Mielke PW. A rapid FORTRAN subroutine for the Fisher exact probability test. *Educ Psychol Meas* 1983, 43:167–171.

139. Berry KJ, Mielke PW. Subroutines for computing exact chi-square and Fisher's exact probability tests. *Educ Psychol Meas* 1985, 45:153–159.

140. Mielke PW, Berry KJ. Asymptotic clarifications, generalizations, and concerns regarding an extended class of matched pairs tests based on powers of ranks. *Psychometrika* 1983, 48:483–485.

141. Mielke PW, Berry KJ. An exact solution to an occupancy problem: a useful alternative to Cochran's *Q* test. *Percept Mot Skills* 1996, 82:91–95.

142. Zimmermann H. Exact calculation of permutational distributions for two dependent samples. *Biom J* 1985, 3:349–352.

143. Zimmermann H. Exact calculation of permutational distributions for two independent samples. *Biom J* 1985, 4:431–434.

144. Mielke PW, Berry KJ, Eighmy JL. A permutation procedure for comparing archaeomagnetic polar directions. In: Eighmy JL, Sternberg RS, eds. *Archaeomagnetic Dating*. Tucson, AZ: University of Arizona Press; 1991, 102–108.

145. Berry KJ, Mielke PW, Kvamme KL. Efficient permutation procedures for analysis of artifact distributions. In: Hietala HJ, ed. *Intrasite Spatial Analysis in Archaeology*. Cambridge: Cambridge University Press; 1984, 54–74.

146. Zimmerman GM, Goetz H, Mielke PW. Use of an improved statistical method for group comparisons to study effects of prairie fire. *Ecology* 1985, 66:606–611.

147. Biondini ME, Mielke PW, Berry KJ. Data-dependent permutation techniques for the analysis of ecological data. *Vegetatio* 1988, 75:161–168.

148. Mielke PW, Berry KJ, Brier GW. Application of multi-response permutation procedures for examining seasonal changes in monthly mean sea-level pressure patterns. *Mon Weather Rev* 1981, 109:120–126.

149. Mielke PW. Meteorological applications of permutation techniques based on distance functions. In: Krishnaiah PR, Sen PK, eds. *Handbook of Statistics*. Vol. 4. Amsterdam: North–Holland; 1984, 813–830.

150. Mielke PW. Geometric concerns pertaining to applications of statistical tests in the atmospheric sciences. *J Atmos Sci* 1985, 42:1209–1212.

151. Wong RKW, Chidambaram N, Mielke PW. Application of multi-response permutation procedures and median regression for covariate analyses of possible weather modification effects on hail responses. *Atmosphere-Ocean* 1983, 21:1–13.

152. Tucker DF, Mielke PW, Reiter ER. The verification of numerical models with multivariate randomized block permutation procedures. *Meteorol Atmos Phys* 1989, 40:181–188.

153. Lee TJ, Pielke RA, Mielke PW. Modeling the clear-sky surface energy budget during FIFE 1987. *J Geophys Res* 1995, 100:25585–25593.

154. Kelly FP, Vonder Haar TH, Mielke PW. Imagery randomized block analysis (IRBA) applied to the verification of cloud edge detectors. *J Atmos Oceanic Technol* 1989, 6:671–679.

155. Cotton WR, Thompson G, Mielke PW. Realtime mesoscale prediction on workstations. *Bull Am Meteorol Soc* 1994, 75:349–362.

156. Mielke PW, Berry KJ, Landsea CW, Gray WM. Artificial skill and validation in meteorological forecasting. *Weather Forecast* 1996, 11:153–169.

157. Mielke PW, Berry KJ, Landsea CW, Gray WM. A single-sample estimate of shrinkage in meteorological forecasting. *Weather Forecast* 1997, 12:847–858.

158. Mehta CR, Patel NR. A hybrid algorithm for Fisher's exact test in unordered $r\times c$ contingency tables. *Commun Stat Theory Methods* 1986, 15:387–403.

159. Mehta CR, Patel NR. FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered $r\times c$ contingency tables (Algorithm 643). *ACM Trans Math Softw* 1986, 12:154–161.

160. Joe H. Extreme probabilities for contingency tables under row and column independence with application to Fisher's exact test. *Commun Stat Theory Methods* 1988, 17:3677–3685.

161. Zar JH. A fast and efficient algorithm for the Fisher exact test. *Behav Res Methods Instrum Comput* 1987, 19:413–414.

162. Romesburg HC, Marshall K, Mauk TP. FITEST: a computer program for "exact chi-square" goodness-of-fit significance tests. *Comput Geosci* 1981, 7:47–58.

163. Mielke PW, Berry KJ. Exact goodness-of-fit probability tests for analyzing categorical data. *Educat Psychol Meas* 1993, 53:707–710.

164. Tritchler DL. An algorithm for exact logistic regression. *J Am Stat Assoc* 1984, 79:709–711.

165. Vollset SE, Hirji KF. A microcomputer program for exact and asymptotic analysis of several 2×2 tables. *Epidemiology* 1991, 2:217–220.

166. Vollset SE, Hirji KF, Elashoff RM. Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *J Am Stat Assoc* 1991, 86:404–409.

167. Berry KJ, Mielke PW. Simulated power comparisons of the asymptotic and nonasymptotic Goodman and Kruskal tau tests for sparse R by C tables. In: Srivastava JN, ed. *Probability and Statistics: Essays in Honor of Franklin A. Graybill*. Amsterdam: North–Holland; 1988, 9–19.

168. Berry KJ, Mielke PW. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educ Psychol Meas* 1988, 48:921–933.

169. Hirji KF, Mehta CR, Patel NR. Computing distributions for exact logistic regression. *J Am Stat Assoc* 1987, 82:1110–1117.

170. Baker RD, Tilbury JB. Rapid computation of the permutation paired and grouped t-tests (Algorithm 283). *Appl Stat* 1993, 42:432–441.

171. Chen RS, Dunlap WP. SAS procedures for approximate randomization tests. *Behav Res Methods Instrum Comput* 1993, 25:406–409.

172. Edgington ES, Khuller PLV. A randomization test computer program for trends in repeated-measures data. *Educ Psychol Meas* 1992, 52:93–95.

173. Mielke PW, Berry KJ. Nonasymptotic inferences based on Cochran's Q test. *Percept Mot Skills* 1995, 81:319–322.

174. Mehta CR, Patel NR, Senchaudhuri P. Exact power and sample-size computations for the Cochran–Armitage trend test. *Biometrics* 1998, 54:1615–1621.

175. Berry KJ, Mielke PW. Computation of exact and approximate probability values for a matched-pairs permutation test. *Commun Stat Simul Comput* 1985, 14:229–248.

176. Berry KJ, Mielke PW. Analysis of multivariate matched-pairs data: a FORTRAN 77 program. *Percept Mot Skills* 1996, 83:788–790.

177. Brockwell PJ, Mielke PW. Asymptotic distributions of matched-pairs permutation statistics based on distance measures. *Aust J Stat* 1984, 26:30–38.

178. Mielke PW, Berry KJ. A extended class of permutation techniques for matched pairs. *Commun Stat Theory Methods* 1982, 11:1197–1207.

179. Mielke PW, Berry KJ, Neidt CO. A permutation test for multivariate matched-pairs analyses: comparisons with Hotelling's multivariate matched-pairs T^2 test. *Psycholo Rep* 1996, 78:1003–1008.

180. Berry KJ, Kvamme KL, Mielke PW. A permutation technique for the spatial analysis of artifacts into classes. *Am Antiq* 1980, 45:55–59.

181. Berry KJ, Kvamme KL, Mielke PW. Improvements in the permutation test for the spatial analysis of the distribution of artifacts into classes. *Am Antiq* 1983, 48:547–553.

182. Berry KJ, Mielke PW. Computation of finite population parameters and approximate probability values for multi-response permutation procedures (MRPP). *Commun Statis Simul Comput* 1983, 12:83–107.

183. Berry KJ, Mielke PW. Computation of exact probability values for multi-response permutation procedures (MRPP). *Commun Stat Simul Comput* 1984, 13:417–432.

184. Berry KJ, Mielke PW. Least absolute regression residuals: analyses of split-plot designs. *Psychol Rep* 1999, 85:445–453.

185. O'Reilly FJ, Mielke PW. Asymptotic normality of MRPP statistics from invariance principles of U-statistics. *Commun Stat Theory Methods* 1980, 9:629–637.

186. Brockwell PJ, Mielke PW, Robinson J. On non-normal invariance principles for multi-response permutation procedures. *Aust J Stat* 1982, 24:33–41.

187. Mielke PW. The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth Sci Rev* 1991, 31:55–71.

188. Mielke PW. Non-metric statistical analyses: some metric alternatives. *J Stat Plann Infer* 1986, 13:377–387.

189. Mielke PW, Berry KJ, Brockwell PJ, Williams JS. A class of nonparametric tests based on multiresponse permutation procedures. *Biometrika* 1981, 68:720–724.

190. Mielke PW, Berry KJ. Multivariate tests for correlated data in completely randomized designs. *J Educ Behav Stat* 1999, 24:109–131.

191. Berry KJ, Mielke PW. Least absolute regression residuals: analyses of block designs. *Psychol Rep* 1998, 83:923–929.

192. Mielke PW, Iyer HK. Permutation techniques for analyzing multi-response data from randomized block experiments. *Commun Stat Theory Methods* 1982, 11:1427–1437.

193. Berry KJ, Mielke PW. A family of multivariate measures of association for nominal independent variables. *Educ Psychol Meas* 1992, 52:41–55.

194. Berry KJ, Mielke PW. A measure of association for nominal independent variables. *Educ Psychol Meas* 1992, 52:895–898.

195. Mielke PW. Goodman–Kruskal tau and gamma. In: Kotz S, Johnson NL, eds. *Encyclopedia of Statistical Sciences*. Vol. III. New York: John Wiley & Sons; 1983, 446–449.

196. Berry KJ, Mielke PW. Nonasymptotic goodness-of-fit tests for categorical data. *Educ Psychol Meas* 1994, 54:676–679.

197. Mielke PW, Berry KJ. Permutation covariate analyses of residuals based on Euclidean distance. *Psychol Rep* 1997, 81:795–802.

198. Mielke PW, Berry KJ. Permutation-based multivariate regression analysis: the case for least sum of absolute deviations regression. *Ann Oper Res* 1997, 74:259–268.

199. Berry KJ, Mielke PW. A FORTRAN program for permutation covariate analyses of residuals based on Euclidean distance. *Psychol Rep* 1998, 82:371–375.

200. Berry KJ, Mielke PW. Least sum of absolute deviations regression: distance, leverage, and influence. *Percept Mot Skills* 1998, 86:1063–1070.

201. Berry KJ, Mielke PW. Least absolute regression residuals: analyses of randomized designs. *Psychol Rep* 1999, 84:947–954.

202. Berry KJ, Mielke PW. A generalized agreement measure. *Educ Psychol Meas* 1990, 50:123–125.

203. Berry KJ, Mielke PW. Agreement measure comparisons between two independent sets of raters. *Educ Psychol Meas* 1997, 57:360–364.

204. Keller-McNulty S, Higgins JJ. Effect of tail weight and outliers and power and type-I error of robust permutation tests for location. *Commun Stat Comput Simul* 1987, 16:17–35.

205. Bailer AJ. Testing variance equality with randomization tests. *J Stat Comput Simul* 1989, 31:1–8.

206. Kim MJ, Nelson CR, Startz R. Mean revision in stock prices? A reappraisal of the empirical evidence. *Rev Econ Stud* 1991, 58:515–528.

207. McQueen G. Long-horizon mean-reverting stock priced revisited. *J Financ Quant Anal* 1992, 27:1–17.

208. Noreen EW. *Computer-Intensive Methods For Testing Hypotheses: An Introduction*. New York: John Wiley & Sons; 1989.

209. Johnston JE, Berry KJ, Mielke PW. Permutation tests: precision in estimating probability values. *Percept Mot Skills* 2007, 105:915–920.

210. Lunneborg CE. *Data Analysis by Resampling: Concepts and Applications*. Pacific Grove, CA: Duxbury; 2000.

211. Good PI. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. 2nd ed. New York: Springer-Verlag; 2000.

212. Good PI. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. 2nd ed. New York: Springer-Verlag; 2000.

213. Good PI. *Resampling Methods: A Practical Guide to Data Analysis*. 2nd ed. Boston, MA: Birkhäuser; 2001.

214. Mielke PW, Berry KJ. *Permutation Methods: A Distance Function Approach*. New York: Springer-Verlag; 2001.

215. Pesarin F. *Multivariate Permutation Tests: With Applications in Biostatistics*. Chichester: John Wiley & Sons; 2001.

216. Lahiri SN. *Resampling Methods for Dependent Data*. New York: Springer-Verlag; 2003.

217. Good PI. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. 3rd ed. New York: Springer-Verlag; 2005.

218. Good PI. *Resampling Methods: A Practical Guide to Data Analysis*. 3rd ed. Boston, MA: Birkhäuser; 2006.

219. Hirji KF. *Exact Analysis of Discrete Data*. Boca Raton, FL: Chapman & Hall/CRC; 2006.

220. Manly BFJ. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC; 2007.

221. Bergman R, Ludbrook J, Spooren WPJM. Different outcomes of the Wilcoxon–Mann–Whitney test from different statistics packages. *Am Stat* 2000, 54:72–77.

222. Hilbe J. Statistical computing software reviews: section editor's notes. *Am Stat* 2000, 54:71.

223. Fisher RA. *Statistical Methods and Scientific Inference*. 2nd ed. New York: Hafner; 1959.

224. Goodman SN, discussants. *p* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993, 137:485–501.

225. Hubbard R. Alphabet soup: blurring the distinctions between *p*'s and α's in psychological research. *Theory Psychol* 2004, 14:295–327.

226. Hubbard R, Bayarri MJ, discussants. Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *Am Stat* 2003, 57:171–182.

