Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 8; Décembre/December 2012

www.jehps.net

The Two-sample Rank-sum Test: Early Development

KENNETH J. BERRY,[1] PAUL W. MIELKE, Jr.,[2] and JANIS E. JOHNSTON[3,4]

Résumé

Nous étudions l'histoire du test statistique de la somme des rangs pour un double échantillon. Alors que la plupart des textes attribuent sa création à Wilcoxon (1945) ou/et à Mann et Whitney (1947), le test a été développé indépendamment par au moins six chercheurs à la fin des années 1940 et au début des années 1950. En complément de Wilcoxon et Mann et Whitney, Festinger (1946), Whitfield (1947), Haldane and Smith (1948), et van der Reyden (1952) ont publié de façon autonome des versions équivalentes, quoique différant par leur méthodologie, du test du rang. Dans cet article, nous décrivons et comparont le développement de ces six approches.

Abstract

The historical development of the two-sample rank-sum test is explored. While most textbooks and journal articles attribute the origin of the two-sample rank-sum test to Wilcoxon (1945) and/or Mann and Whitney (1947), the test was independently developed by at least six researchers in the late 1940s and early 1950s. In addition to Wilcoxon and Mann and Whitney, Festinger (1946), Whitfield (1947), Haldane and Smith (1948), and van der Reyden (1952) autonomously published methodologically distinct, but equivalent versions of the two-sample rank-sum test. In this article the historical development of the six approaches are described and compared.

Key names: L. Festinger, J. B. S. Haldane, H. B. Mann, D. van der Reyden, C. A. B. Smith, J. W. Whitfield, D. R. Whitney, F. Wilcoxon

Keywords: Combinatoric Methods, History of Statistics, Partitions, Two-sample Rank-sum Test

---

[1]Department of Sociology, Colorado State University, Fort Collins, CO 80523–1784, USA. e-mail: berry@mail.colostate.edu.

[2]Department of Statistics, Colorado State University, Fort Collins, CO 80523–1877, USA.

[3]Food and Nutrition Service, United States Department of Agriculture, Alexandria, VA 22302–1500, USA.

[4]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.

1 Introduction

In the 1930s and 1940s it was widely recognized that when extreme values (outliers) were present in sets of data, the assumption of normality underlying conventional parametric tests such as t tests, analysis of variance, and correlation was untenable and the results were therefore questionable (cf. Vankeerberghen, Vandenbosch, Smeyers-Verbeke, & Massart, 1991). In response, a number of non-parametric distribution-free counterparts to existing parametric tests were developed, typically based on converting raw measurements to rank values, notwithstanding the potential loss of information in reducing numerical values to ranks. Rank tests such as Kendall's measure of rank correlation (Kendall, 1938), Friedman's two-way analysis of variance by ranks (Friedman, 1937), and the Kruskal–Wallis one-way analysis of variance by ranks (Kruskal & Wallis, 1952) are very robust to outliers (Potvin & Roff, 1993). The non-parametric counterpart to the two-sample t test is the two-sample rank-sum test based on the sums of the ranks in the two samples, independently developed by Wilcoxon (1945), Festinger (1946), Mann and Whitney (1947), Whitfield (1947), Haldane and Smith (1948), and van der Reyden (1952).

The logic underlying the two-sample rank-sum test is straightforward. The data consist of two independent samples drawn from identically distributed populations. Let $x_1, x_2, \ldots, x_n$ denote the first random sample of size n and let $y_1, y_2, \ldots, y_m$ denote the second random sample of size m. Assign the ranks 1 to $n + m$ to the combined observations from smallest to largest without regard to sample membership and let R_k denote the rank assigned to the $n + m$ observations for $k = 1, \ldots, n + m$. Let T_x and T_y denote the sums of the ranks from the first and second samples, respectively, and let $T = T_x$. Finally, note that

$$T_x + T_y = \frac{(n+m)(n+m+1)}{2} \, .$$

The null hypothesis simply states that each of the possible arrangements of the $n + m$ observations to the two samples with n values in the first sample and m values in the second sample occurs with equal probability. The exact lower (upper) one-sided probability value of an observed value of T, T_o, is the proportion of all possible $\binom{n+m}{n}$ T values less (greater) than or equal to T_o.

For example, consider $n + m = 10$ graduate students in a graduate seminar with $n = 5$ males and $m = 5$ females. Fig. 1 contains the raw data where x_i and y_j represent current ages in years. Fig. 2 lists the numerical ages, sex (M, F), and associated ranks of the $n + m = 10$

Males:	20,	22,	23,	28,	32
Females:	24,	25,	27,	30,	45

Fig. 1: Current ages of $n = 5$ male and $m = 5$ female graduate students.

graduate students based on the data in Fig. 1. For the data in Fig. 2, $T_\text{o} = T_x = 22$, $T_y = 33$, and only 39 (224) of the

$$\binom{n+m}{n} = \frac{(n+m)!}{n!\,m!} = \frac{(5+5)!}{5!\,5!} = 252$$

Age:	20,	22,	23,	24,	25,	27,	28,	30,	32,	45
Sex:	M,	M,	M,	F,	F,	F,	M,	F,	M,	F
Rank:	1,	2,	3,	4,	5,	6,	7,	8,	9,	10

Fig. 2: Age, sex, and corresponding rank of $n = 5$ male and $m = 5$ female graduate students.

Table 1: Listing of the 39 combinations of ranks with T values less than or equal to $T_o = 22$.

Number	Ranks	T	Number	Ranks	T	Number	Ranks	T
1	1 2 3 4 5	15	14	1 2 3 5 9	20	27	1 3 4 6 7	21
2	1 2 3 4 6	16	15	1 2 3 6 8	20	28	2 3 4 5 7	21
3	1 2 3 4 7	17	16	1 2 4 5 8	20	29	1 2 3 6 10	22
4	1 2 3 5 6	17	17	1 2 4 6 7	20	30	1 2 3 7 9	22
5	1 2 3 4 8	18	18	1 3 4 5 7	20	31	1 2 4 5 10	22
6	1 2 3 5 7	18	19	2 3 4 5 6	20	32	1 2 4 6 9	22
7	1 2 4 5 6	18	20	1 2 3 5 10	21	33	1 2 4 7 8	22
8	1 2 3 4 9	19	21	1 2 3 6 9	21	34	1 2 5 6 8	22
9	1 2 3 5 8	19	22	1 2 3 7 8	21	35	1 3 4 5 9	22
10	1 2 3 6 7	19	23	1 2 4 5 9	21	36	1 3 4 6 8	22
11	1 2 4 5 7	19	24	1 2 4 6 8	21	37	1 3 5 6 7	22
12	1 3 4 5 6	19	25	1 2 5 6 7	21	38	2 3 4 5 8	22
13	1 2 3 4 10	20	26	1 3 4 5 8	21	39	2 3 4 5 7	22

possible values of T are less (greater) than or equal to $T_o = 22$. For clarity, the 39 T values less than or equal to $T_o = 22$ are listed in Table 1. Thus, for the data in Figs. 1 and 2, the exact lower (upper) one-sided probability of $T_o = 22$ is $39/252 \doteq 0.1548$ ($224/252 \doteq 0.8889$).

2 Development of the Two-sample Rank-sum Test

Stigler's law of eponymy states that "[n]o scientific discovery is named after its original discoverer" (Stigler, 1999, p. 277). Stigler observed that names are not given to scientific discoveries or inventions by historians of science, but by the community of practicing scientists, most of whom have no historical expertise. Moreover, the award of an eponym must be made on the basis of the scientific merit of originality and not upon personal friendship, national affiliation, or political pressure (Stigler, 1999, pp. 280–281).

The two-sample rank-sum test, commonly referenced with the eponym "Wilcoxon–Mann–Whitney test," was invented and reinvented by a number of researchers in the late 1940s and early 1950s (considering co-authors as a single contributor): F. Wilcoxon in 1945, L. Festinger in 1946, H. B. Mann and D. R. Whitney in 1947, J. W. Whitfield in 1947, J. B. S. Haldane and C. A. B. Smith in 1948, and D. van der Reyden in 1952. Because Festinger (1946) published his version of the two-sample rank-sum test in *Psychometrika*, Haldane and Smith (1948) published their version in *The Annals of Eugenics*, and van der Reyden (1952) published his version in *The Rhodesia Agricultural Journal*, their work was largely overlooked by statisticians at the time. Although Whitfield (1947) published his version in *Biometrika*, which was widely read by statisticians, the article contained no references to earlier work on the two-sample rank-sum test and was, in fact, simply an examination of rank correlation between two variables, one dichotomous and one ranked, utilizing a pairwise procedure previously developed by M. G. Kendall (1938).

The development of the two-sample rank-sum test has never been adequately documented. In 1957 Kruskal (1957) published a short note titled "Historical notes on the Wilcoxon unpaired two-sample test" that discussed early contributions by Deuchler (1914), Lipmann (1908), Whitfield (1947), Gini (1916/1959), and Ottaviani (1939), most of which appeared many years prior to Wilcoxon's seminal article in 1945. What is missing from the literature is a detailed treatment of the six major developers of the two-sample rank-sum test, each of which used a different approach to generate exact frequency distributions, and four of which have been largely ignored in the literature. The six contributors are Wilcoxon (1945), Festinger (1946), Mann and Whitney (1947), Whitfield (1947), Haldane and Smith (1948), and van der Reyden (1952), and the neglected four are, of course, Festinger, Haldane and Smith, Whitfield, and van der Reyden.

The history of the development of the two-sample rank-sum test provides an opportunity to give the six contributors their due recognition. Prior to the age of computers, researchers relied on innovative and, oftentimes, very clever methods for computation. Each of these six investigators developed very different alternative computational methods that are interesting both from a statistical and a historical perspective. A common notation, detailed descriptions, and example analyses will hopefully provide researchers with an appreciation of the rich, but neglected, history of the development of the two-sample rank-sum test.

3 Wilcoxon's Rank-sum Test

Frank Wilcoxon earned a B.Sc. degree from Pennsylvania Military College in 1917, a M.S. degree in chemistry from Rutgers University in 1921, and a Ph.D. in chemistry from Cornell University in 1924. Wilcoxon spent most of his life as a chemist working for the Boyce Thompson Institute for Plant Research, the Atlas Powder Company, and the American Cyanamid Company. While at the Boyce Thompson Institute, Wilcoxon, together with chemist William John (Jack) Youden and biologist F. E. Denny led a group in studying Fisher's newly published *Statistical Methods for Research Workers* (Fisher, 1925). This introduction to statistics had a profound effect on the subsequent careers of Wilcoxon and Youden as both became leading statisticians of the time. Wilcoxon retired from the American Cyanamid Company in 1957 and moved to Florida. Three years later, at the invitation of Ralph Bradley, Wilcoxon joined the faculty at Florida State University where he helped to develop the Department of Statistics. As Bradley related, he and Wilcoxon had met several times at Gordon Research Conferences, and in 1959 Bradley was recruited from Virginia Polytechnic Institute to create and head a new department of statistics at Florida State University in Tallahassee. Since Wilcoxon was living in Florida, Bradley persuaded Wilcoxon to come out of retirement and join the department. Wilcoxon agreed to a half-time position as he wanted time off to kayak and ride his motorcycle (Hollander, 2000). Wilcoxon died on 18 November 1965 after a brief illness at the age of 73 (Bradley, 1966, 1997; Bradley & Hollander, 2001).

In 1945 Wilcoxon introduced a two-sample test statistic, W, for rank order statistics.[1] The stated purpose was to develop methods in which ranks $1, 2, 3, \ldots$ are substituted for the actual numerical values in order to obtain a rapid approximation of the significance of the differences in two-sample paired and unpaired experiments (Wilcoxon, 1945, p. 80). In this very brief paper of only three pages, Wilcoxon considered the case of two samples of equal sizes and provided a table of exact probabilities for values of the lesser of the two sums of ranks for both two-sample paired and two-sample unpaired experiments. In the case of the two-sample unpaired test, a table provided exact probability values for 5 to 10 replicates in each sample. Bradley has referred to the unpaired and paired rank tests as the catalysts for the flourishing of non-parametric statistics (Hollander, 2000, p. 88).

Wilcoxon (1945) showed that in the case of unpaired samples with rank numbers from 1 to $2q$, where q denotes the number of ranks (replicates) in each sample, the minimum sum of ranks possible is given by $q(q+1)/2$, where W is the sum of ranks in one sample, continuing by steps up to the maximum sum of ranks given by $q(3q+1)/2$. For example, consider two samples of $q = 5$ measurements ranked from 1 to $2q = 10$. The minimum sum of ranks for either group is $\{1+2+3+4+5\} = 5(5+1)/2 = 15$ and the maximum sum of ranks is $\{6+7+8+9+10\} = 5[(3)(5)+1]/2 = 40$. Wilcoxon explained that these two values could be obtained in only one way, but intermediate sums could be obtained in more than one way. For example, the sum of $T = 20$ could be obtained in $q = 5$-part seven ways, with no part greater than $2q = 10$: $\{1, 2, 3, 4, 10\}$, $\{1, 2, 3, 5, 9\}$, $\{1, 2, 3, 6, 8\}$, $\{1, 2, 4, 5, 8\}$, $\{1, 2, 4, 6, 7\}$, $\{1, 3, 4, 5, 7\}$, and $\{2, 3, 4, 5, 6\}$. The number of ways each sum could arise is given by the number of q-part, here 5-part, partitions of $T = 20$, the sum in question.[2]

This was not a trivial problem to solve, as calculating the number of partitions is quite difficult, even today with the availability of high-speed computing. In general, the problem is known as the "subset-sum problem" and requires a generating function to solve. The difficulty is in finding all subsets of a set of distinct numbers that sum to a specified total. The approach that Wilcoxon took was ingenious and is worth examining, as the technique became the basic method for other researchers as well as the basis for several computer algorithms in later years. Wilcoxon showed that the required partitions were equinumerous with another set of partitions, r, that were much easier to enumerate. He defined r as the serial number of T in the possible series of sums, beginning with 0, i.e., 0, 1, $2, \ldots, r$. This was a technique that Wilcoxon apparently came across while reading a book on combinatorial analysis by Percy Alexander MacMahon (MacMahon, 1916). MacMahon's monumental two-volume work on *Combinatory Analysis*, published in 1916, contained a section in Volume II, Chapter III, on "Ramanujan's Identities" in which MacMahon demonstrated the relationship between the number of q-part unequal partitions without repetitions with no part greater than $2q$ and the number of partitions with repetitions with no part greater than q (MacMahon, 1916, pp. 33–48).

For example, consider as previously, $q = 5$ replications of measurements on two samples

[1]Most textbooks and articles use the letter W to indicate the Wilcoxon two-sample rank-sum test statistic. Actually, Wilcoxon never used W in either his 1945 or 1947 articles; he always used T to indicate the total (sum) of the ranks.

[2]Wilcoxon's use of the term "partitions" here is a little misleading. These are actually sums of $T = 20$, each sum consisting of five integer values between 1 and $2q = 10$ with no integer value repeated e.g., $\{1, 2, 3, 4, 10\} = 20$ which consists of five non-repeating integer values, but not $\{5, 7, 8\} = 20$ which consists of only three integer values, nor $\{1, 3, 3, 5, 8\} = 20$ which contains multiple values of 3.

Table 2: Illustrative table comparing the $q = 5$-part partitions of $T = 20$ with the corresponding partitions of $r = 5$.

	Partition									
Number	$q = 5,\ T = 20$					$r = 5$				
1	1	2	3	4	10	1	1	1	1	1
2	1	2	3	5	9		1	1	1	2
3	1	2	3	6	8			1	2	2
4	1	2	4	5	8			1	1	3
5	1	2	4	6	7				2	3
6	1	3	4	5	7				1	4
7	2	3	4	5	6					5

and assign ranks 1 through $2q = 10$ to the data: $\{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\}$. Recall that the lowest possible sum is 15 and the highest possible sum is 40. Then the question is: In how many ways can a total of $T = 20$ be obtained, i.e., how many unequal $q = 5$-part partitions of $T = 20$ exist, having no part greater than 10 and no repetition of values? As shown in Table 2, there are seven such partitions. Now, 20 is sixth in the possible series of totals, as shown in Fig. 3. Therefore, $r = 5$ and the total number of partitions that sum to $T = 20$ is equiv-

T:	15,	16,	17,	18,	19,	20,	21,	22,	23,	$\ldots$, 40
r:	0,	1,	2,	3,	4,	5,	6,	7,	8,	$\ldots$, 25

Fig. 3: Values of T up to and including $T = 23$ with corresponding values of r.

alent to the total number of partitions with repetitions that sum to 5 with no part greater than $q = 5$; specifically, $\{5\}$, $\{1,\ 4\}$, $\{2,\ 3\}$, $\{1,\ 1,\ 3\}$, $\{1,\ 2,\ 2\}$, $\{1,\ 1,\ 1,\ 2\}$, and $\{1,\ 1,\ 1,\ 1,\ 1\}$.[3] Wilcoxon capitalized on the relationship between the subset-sum problem with $T = 20$ and the partition problem with $r = 5$, to enumerate the partitions of $r = 5$ from an available table of partitions by Whitworth (1942), which then corresponded to the more difficult enumeration of the 5-part partitions of $T = 20$.

The exact lower one-sided probability (P) value of $T = 20$ is given by

$$P = \left\{ 1 + \sum_{i=1}^{r} \sum_{j=1}^{q} \mathbb{P}_j^i - \sum_{k=1}^{r-q} \left[(r - q - k + 1) \mathbb{P}_{q-1}^{q-2+k} \right] \right\} \bigg/ \frac{(2q)!}{(q!)^2} \,, \tag{3.1}$$

where $\mathbb{P}_j^i$ represents the number of j-part partitions of i; r is the serial number of possible rank totals, 0, 1, 2, $\ldots$, r; and q is the number of replicates (Wilcoxon, 1945, p. 82). If $q \geq r$, the

[3]These are, of course, true partitions consisting of one to five integer values between 1 and 5, summing to 5 with repetitions allowed.

summation $\sum_{k=1}^{r-q}$ is assumed to be zero. For the example data, Eq. 3.1 is

$$P = \left\{ 1 + \sum_{i=1}^{5} \sum_{j=1}^{5} \mathbb{P}_j^i - \sum_{k=1}^{5-5} \left[(5-5-k+1)\mathbb{P}_{5-1}^{5-2+k} \right] \right\} \Big/ \frac{10!}{(5!)^2} \, ,$$

and the exact lower one-sided probability value is

$$P = \{1+1+2+3+5+7-0\} \, / \left[3{,}628{,}800/(120)^2 \right] = 19/252 = 0.0754.$$

The equivalence between the number of unequal q-part partitions of T with no part greater than $2q$ and the number of partitions of r with no part greater than q used by Wilcoxon greatly reduced the calculations required.

4 Festinger's Rank-sum Test

Leon Festinger earned his B.Sc. degree in psychology from City College of New York in 1939, then moved to the University of Iowa to earn his Ph.D. in psychology in 1942 under the renowned social psychologist Kurt Lewin. Festinger is best known for his work in social psychology and, especially, his theory of cognitive dissonance, but he was also an accomplished statistician, working in the area of non-parametric statistics. After earning his Ph.D., Festinger worked as a Research Associate at the University of Iowa, then joined the University of Rochester as a senior statistician in 1943. In 1945, Festinger moved to the Massachusetts Institute of Technology, the University of Michigan in 1948, the University of Minnesota in 1951, Stanford University in 1955, and finally to the New School for Social Research (now, New School University) in 1968 where he remained until his death from liver cancer on 11 February 1989 at the age of 69 (Moscovici, 1989).

In 1946 Festinger introduced a statistical test of differences between two independent means by first converting raw scores to ranks, then testing the difference between the means of the ranks (Festinger, 1946). The stated purpose of the new test was the need for a statistic that could be applied without making any assumption concerning the distribution function in the parent population (Festinger, 1946, p. 97). Festinger provided tables for tests of significance based on exact probabilities for the 0.05 and 0.01 confidence levels for $n = 2, \ldots, 15$, the smaller of the two samples, and $m = 2, \ldots, 38$, the larger sample. Apparently, Festinger's solution to the two-sample rank problem was developed independently of Wilcoxon's solution; moreover, Festinger's tables considered both equal and unequal sample sizes, whereas Wilcoxon's method allowed for only $m = n$ (Wilcoxon, 1945). In addition, the approach that Festinger took was quite different from that of Wilcoxon. While both approaches generated all possible permutations of outcomes, Festinger's procedure was considerably simpler to implement and was based on a unique and ingenious recursive generation method.

Consider two independent samples $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ with $n \leq m$. Combining the samples x and y and assigning ranks to each case from 1 to $m + n$ structures the question as to the probability of obtaining any specified difference between sample ranks if both samples are drawn at random from the same population. Stated in terms of sums of ranks: What is the probability of obtaining any specified sum of ranks of n cases selected at random from the total of $m+n$ cases? The problem for Festinger was to generate exact probability distributions for sums of ranks for specified values of m and n.

For simplicity, consider first $m = 2$ and $n = 2$. The possible combinations of $m + n = 2 + 2 = 4$ considered $n = 2$ at a time are $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, and $\{3, 4\}$, yielding sums of 3, 4, 5, 5, 6, and 7, respectively. Thus, the frequency distribution of the sums is 3(1), 4(1), 5(2), 6(1), and 7(1), where the frequencies are enclosed in parentheses. If each case is independent of every other case and equally likely to be drawn, then each combination is equiprobable. However, as Festinger showed, there is an alternative way to generate this frequency distribution of sums. The frequency distribution of sums for $\binom{m+n}{n}$ can be constructed from the frequency distributions of sums for $\binom{m+n-1}{n}$ and $\binom{m+n-1}{n-1}$, as illustrated in Table 3.[4] The frequency distribution of $\binom{m+n-1}{n} = \binom{2+2-1}{2} = \binom{3}{2}$ is listed in Column 1 of Table 3 and the frequency distribution of sums for $\binom{m+n-1}{n-1} = \binom{2+2-1}{2-1} = \binom{3}{1}$ is listed in Column 2 of Table 3. Note that the frequency distribution of sums for $\binom{3}{1}$ is offset from the frequency distribution of sums for $\binom{3}{2}$. Since the sum of ranks below the value 5 would not be affected by the addition of a 4th case to the ranks of $\binom{3}{2}$, only the totals of 5, 6, and 7 would be augmented by one or more possibilities. In general, the starting value for frequency distribution $\binom{m+n-1}{n-1}$ is given by $n(n + 1)/2 + m$; in this case, $2(2 + 1)/2 + 2 = 5$. Thus, the frequency distribution of sums for $\binom{m+n}{n} = \binom{4}{2}$ in Column 3 is constructed from the frequency distributions of sums for $\binom{m+n-1}{n} = \binom{3}{2}$ and $\binom{m+n-1}{n-1} = \binom{3}{1}$ in Columns 1 and 2 in Table 3, respectively, by simply adding across Columns 1 and 2 to obtain the frequency distribution of sums for $\binom{4}{2}$ in Column 3.

Once the exact frequency distributions of sums for $m+n$ ranks considered $n = 2$ at a time are established, it is relatively straightforward to construct exact frequency distributions of sums for $m + n$ ranks considered $n = 3$ at a time, using the same approach. This method allowed Festinger to recursively generate exact frequency distributions of sums for any combination of $m + n$ and n.

Finally, Festinger proposed a convenient alternative for summarizing and presenting the frequency distributions of sums. He replaced the sums of ranks of the smaller of the two samples with the absolute deviation (d) of the mean of the ranks of the smaller sample from the mean of the ranks of the total group, using

$$d = \left| \frac{\sum_{i=1}^{n} R_i}{n} - \frac{m+n+1}{2} \right| , \tag{4.1}$$

where n is the number of cases in the smaller sample, $m + n$ is the number of cases in both samples combined, and $\sum_{i=1}^{n} R_i$ is the sum of the ranks of the cases in the smaller of the two samples. The last term in Eq. 4.1 is, of course, the mean of the $m + n$ ranks. Festinger then presented two tables containing the d values necessary for tests of significance at the 0.01 and 0.05 levels of confidence. For values of n from 2 to 12, the Festinger tables listed values of d from $m = 2$ to $m = 38$.

[4]The decomposition $\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$ has been well known since Blaise Pascal's *Traité du triangle arithmétique* was published in 1665, three years after his death (Pascal, 1665/1959). Thus, considering any one of n objects, $\binom{n-1}{r-1}$ gives the number of combinations that include it and $\binom{n-1}{r}$ the number of combinations that exclude it.

Table 3: Generation of frequency arrays for 3, 4, 5, 6, and 7 objects considered $n = 2$ at a time.

	Column								
	1	2	3	4	5	6	7	8	9
Sum	$\binom{3}{2}$	$\binom{3}{1}$	$\binom{4}{2}$	$\binom{4}{1}$	$\binom{5}{2}$	$\binom{5}{1}$	$\binom{6}{2}$	$\binom{6}{1}$	$\binom{7}{2}$
3	1		1		1		1		1
4	1		1		1		1		1
5	1	1	2		2		2		2
6		1	1	1	2		2		2
7		1	1	1	2	1	3		3
8				1	1	1	2	1	3
9				1	1	1	2	1	3
10						1	1	1	2
11						1	1	1	2
12								1	1
13								1	1

5 Mann and Whitney's Rank-sum Test

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

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

While Mann was at The Ohio State University, one of his graduate students was Donald Ransom Whitney. Whitney had earned his B.A. degree in mathematics from Oberlin College in 1936 and his M.S. degree in mathematics from Princeton University in 1939. After service in the Navy during World War II, Whitney enrolled in the Ph.D. program at The Ohio State University in 1946, where eventually he came to work under Henry Mann. After receiving his Ph.D. in mathematics in 1949, Whitney remained at The Ohio State University, eventually becoming Chair of the newly established Department of Statistics in 1974. Whitney retired from The Ohio State University in 1982, whereupon he received the University Distinguished Service Award. Whitney died on 16 August 2007 at the age of 92 (Willke, 2008).

In 1947 Mann and Whitney, acknowledging the previous work by Wilcoxon on the two-sample rank sum test (Wilcoxon, 1945), proposed an equivalent test statistic, U, based on the relative ranks of two samples denoted by $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ and for which they computed exact probability values (Mann & Whitney, 1947). Like Festinger, Mann and Whitney utilized a recurrence relation involving n and m and, using this relation, computed tables of exact probability values for U up to $n = m = 8$, many more, they noted, than the few probability values provided by Wilcoxon. As Mann and Whitney explained, let the measurements $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ be arranged in order and let U count the number of times a y precedes an x. For example, given $n = 4$ x values and $m = 2$ y values, consider the sequence $\{x, y, x, x, y, x\}$ where $U = 4$: the first y precedes three x values and the second y precedes one x value; thus, $U = 3 + 1 = 4$. Also, let the Wilcoxon statistic, W, be the sum of the m rank order statistics $\{y_1, y_2, \ldots, y_m\}$. The relationship between Wilcoxon's W statistic and Mann and Whitney's U statistic can be expressed as

$$U = mn + \frac{m(m+1)}{2} - W$$

and $0 \leq U \leq mn$. Mann and Whitney noted that since Wilcoxon only considered the case of $n = m$, it seemed worthwhile to extend this important work to $n \neq m$ and larger values of n and m, apparently unaware of the 1946 article by Festinger who also considered $n \neq m$.

Consider again the ordered sequences of n x and m y values, replace each x with a 0 and each y with a 1, let U denote the number of times a 1 precedes a 0, and let $\bar{p}_{n,m}(U)$ represent the number of sequences of n 0s and m 1s in each of which a 1 precedes a 0 U times. For example, suppose the sequence is $\{1, 1, 0, 0, 1, 0\}$, then $U = 7$ as the first 1 precedes three 0 values, the second 1 precedes the same three 0 values, and the third 1 precedes only one 0 value. Mann and Whitney then developed the recurrence relation:

$$\bar{p}_{n,m}(U) = \bar{p}_{n-1,m}(U - m) + \bar{p}_{n,m-1}(U) \ , \tag{5.1}$$

where $\bar{p}_{n-1,m}(U - m) = 0$ if $U \leq m$.

An example of the recurrence relation illustrates the Mann–Whitney procedure. Table 4 lists all the sequences of 0s and 1s and corresponding values of U for $\bar{p}_{n,m}(U)$, $\bar{p}_{n-1,m}(U - m)$, and $\bar{p}_{n,m-1}(U)$ for $n = 4$ and $m = 2$. There are $\binom{m+n}{m} = \binom{2+4}{2} = 15$ values of U in the first sequence of 0s and 1s in Table 4, $\binom{m+n-1}{m} = \binom{2+4-1}{2} = 10$ values of U in the second sequence of 0s and 1s, and $\binom{m-1+n}{m-1} = \binom{2-1+4}{2-1} = 5$ values of U in the third sequence of 0s and 1s.[5] To illustrate the recurrence procedure with $U = 3$, $\bar{p}_{n,m}(3) = 2$, as there are two occurrences of $U = 3$ (in Rows 4 and 7) in the leftmost column of sequences in Table 4. Then, $\bar{p}_{n-1,m}(U - m) = \bar{p}_{4-1,2}(3 - 2) = 1$, as there is only a single occurrence of $U = 1$ (in Row 2) in the middle column of sequences in Table 4, and $\bar{p}_{n,m-1}(U) = \bar{p}_{4,2-1}(3) = 1$, as there is only a single occurrence of $U = 3$ (in Row 4) in the rightmost column of sequences in Table 4. Then, following Eq. 5.1, $2 = 1 + 1$.

Given that under the null hypothesis each of the $(n + m)!/(n!\ m!)$ sequences of n 0s and m 1s is equally likely, let $p_{n,m}(U)$ represent the probability of a sequence in which a 1 precedes a 0 U times. For example, for $U = 3$ in Table 4,

$$p_{n,m}(U) \times \frac{n!\ m!}{(n+m)!} = p_{4,2}(3) \times \frac{4!\ 2!}{(4+2)!} = \frac{2}{15} = 0.1333 \ .$$

[5] Here, the decomposition is identical to that of Festinger (1946).

Table 4: Sequences of $n = 4$ 0s and $m = 2$ 1s for $\bar{p}_{n,m}(U)$, $\bar{p}_{n-1,m}(U-m)$, and $\bar{p}_{n,m-1}(U)$.

	$\bar{p}_{n,m}(U)$		$\bar{p}_{n-1,m}(U-m)$		$\bar{p}_{n,m-1}(U)$	
Row	Sequence	U	Sequence	U	Sequence	U
1	0 0 0 0 1 1	0	0 0 0 1 1	0	0 0 0 0 1	0
2	0 0 0 1 0 1	1	0 0 1 0 1	1	0 0 0 1 0	1
3	0 0 1 0 0 1	2	0 1 0 0 1	2	0 0 1 0 0	2
4	0 1 0 0 0 1	3	1 0 0 0 1	3	0 1 0 0 0	3
5	1 0 0 0 0 1	4	0 0 1 1 0	2	1 0 0 0 0	4
6	0 0 0 1 1 0	2	0 1 0 1 0	3		
7	0 0 1 0 1 0	3	1 0 0 1 0	4		
8	0 1 0 0 1 0	4	0 1 1 0 0	4		
9	1 0 0 0 1 0	5	1 0 1 0 0	5		
10	0 0 1 1 0 0	4	1 1 0 0 0	6		
11	0 1 0 1 0 0	5				
12	1 0 0 1 0 0	6				
13	0 1 1 0 0 0	6				
14	1 0 1 0 0 0	7				
15	1 1 0 0 0 0	8				

Mann and Whitney also provided a recurrence relation for the probability values of U given by

$$p_{n,m}(U) = \frac{n}{n+m} p_{n-1,m}(U-m) + \frac{m}{n+m} p_{n,m-1}(U) \;,$$

where

$$p_{n-1,m}(U-m) = \bar{p}_{n-1,m}(U-m) \times \frac{(n-1)!\; m!}{(n+m-1)!}$$

and

$$p_{n,m-1}(U) = \bar{p}_{n,m-1}(U) \times \frac{n!\; (m-1)!}{(n+m-1)!} \;.$$

Thus, for $U = 3$ in Table 4,

$$\begin{aligned} p_{4,2}(3) &= \frac{4}{4+2} p_{4-1,2}(3-2) + \frac{2}{4+2} p_{4,2-1}(3) \\ \frac{2}{15} &= \left(\frac{4}{6}\right)\left(\frac{1}{10}\right) + \left(\frac{2}{6}\right)\left(\frac{1}{5}\right) \\ \frac{2}{15} &= \frac{1}{15} + \frac{1}{15} \;. \end{aligned}$$

Mann and Whitney used this recurrence relation to construct tables of exact probability values up to and including $n = m = 8$. Finally, from the recurrence relation Mann and Whitney derived explicit expressions for the mean, variance, and various higher moments for U, and noted that the limit of the distribution is normal if $\min(n, m) \to \infty$ (Mann & Whitney, 1947).

It should be noted that in 1914 Gustav Deuchler suggested an approach that was essentially the same as that used by Mann and Whitney in their treatment of the two-sample rank sum test (Deuchler, 1914). Deuchler's work in this area seems to have been neglected, but W. H. Kruskal attempted to redress this neglect in a 1957 article on "Historical notes on the Wilcoxon unpaired two-sample test" in the *Journal of the American Statistical Association* (Kruskal, 1957). In their 1952 article W. H. Kruskal and W. A. Wallis provided a list of independent discoveries of the Wilcoxon two-sample test (Kruskal & Wallis, 1952) and this 1957 article is, in part, an attempt to update that list. Also mentioned in the 1957 article, but omitted in the 1952 article, was a 1947 article by J. W. Whitfield (1947) who essentially independently discovered the Mann–Whitney test.

6 Whitfield's Rank-sum Test

Little is known about John W. Whitfield other than that for most of his career he was attached to the Applied Psychology Research Unit (Medical Research Council) at the University of Cambridge. In 1947 Whitfield proposed a measure of rank correlation between two variables, one of which was ranked and the other dichotomous (Whitfield, 1947). While not presented as a rank-sum test *per se*, the article by Whitfield is of historical importance as it is occasionally cited as an independent discovery of the two-sample rank-sum test (e.g., Kruskal, 1957, pp. 358–359).

Whitfield considered the dichotomous variable as a ranking composed entirely of two sets of tied rankings. An example will illustrate the procedure. Following Whitfield, consider the ranked data in Fig. 4, where the $-$ and $+$ signs indicate the dichotomous variable and the ranks are from 1 to 6. Let $m = 2$ denote the number of ranks in the $+$ group and let $n = 4$

1	2	3	4	5	6
$-$	$+$	$-$	$-$	$-$	$+$

Fig. 4: Rankings of a dichotomous variable.

denote the number of ranks in the $-$ group.

Now consider the $n = 4$ ranks in the group identified by a $-$ sign: 1, 3, 4, and 5. Beginning with rank 1, in this case with a $-$ sign, there are no ranks with a $+$ sign to the left of rank 1 and two ranks with a $+$ sign to the right of rank 1 (ranks 2 and 6); so compute $0 - 2 = -2$. For rank 3 with a $-$ sign, there is one rank to the left of rank 3 with a $+$ sign (rank 2) and one rank to the right of rank 3 with a $+$ sign (rank 6); so compute $1 - 1 = 0$. For rank 4 with a $-$ sign, there is one rank to the left of rank 4 with a $+$ sign (rank 2) and one rank to the right of rank 4 with a $+$ sign (rank 6); so compute $1 - 1 = 0$. Finally, for rank 5 with a $-$ sign, there is one rank to the left of rank 5 with a $+$ sign (rank 2) and one rank to the right of rank 5 with a $+$ sign (rank 6); so compute $1 - 1 = 0$. The sum of the differences is $S = -2 + 0 + 0 + 0 = -2$. In this manner, Whitfield's approach incorporated unequal sample sizes with $m \neq n$ as well as tied ranks.

Since the number of possible pairs of $m+n$ consecutive integers is given by $(m+n)(m+n-1)/2$, Whitfield defined and calculated his test statistic as

$$\tau = \frac{2S}{(m+n)(m+n-1)} = \frac{2(-2)}{(2+4)(2+4-1)} = \frac{-4}{30} = -0.1333 \; .$$

Whitfield's S is directly related to the U statistic of Mann and Whitney (1947) and, hence, to the W statistic of Wilcoxon (1945).[6] Compare statistic S with the U statistic of Mann and Whitney. There are $m = 2$ $+$ signs and $n = 4$ $-$ signs in Fig. 4, so considering the lesser of the two (the $m = 2$ $+$ signs), the first $+$ sign (rank 2) precedes three $-$ signs (ranks 3, 4, and 5) and the second $+$ sign precedes no $-$ signs, so $U = 3 + 0 = 3$. The relationship between Whitfields's S and Mann and Whitney's U is given by $S = 2U - mn$ (Jonckheere, 1954; Kruskal, 1957); thus, $S = 2(3) - (2)(4) = 6 - 8 = -2$. For the example data in Fig. 4, the Wilcoxon test statistic for the smaller of the two sums (with the $m = 2$ $+$ signs) is $W = 2 + 6 = 8$ and the relationship with S is given by $S = m(m+n+1) - 2W$; thus, $S = 2(2+4+1) - (2)(8) = 14 - 16 = -2$.

As Whitfield noted, the calculation of S was fashioned after a procedure first introduced by Kendall in 1945[7] and Whitfield was apparently unaware of the work by Wilcoxon, Festinger, and Mann and Whitney, as they are not referenced in the Whitfield article. Kendall (1945) considered the number of concordant (C) and discordant (D) pairs, of which there are a total of $(m + n)(m + n - 1)/2$ pairs when there are no ties in the $m + n$ consecutive integers. For the example data in Fig. 4 there are $(2 + 4)(2 + 4 - 1)/2 = 15$ pairs. Table 5 numbers and lists the 15 pairs, the concordant/discordant classification of pairs, and the pair values, where concordant pairs $(-, -$ and $+, +)$ are given a value of 0, and discordant pairs $(+, -$ and $-, +)$ are given values of $+1$ and -1, respectively. The sum of the pair values in Table 5 for the 15 pairs is $S = -5 + 3 = -2$.

Today it is well known, although poorly documented, that when one classification is a dichotomy and the other is ordered, with or without tied values, the S test of Kendall is equivalent to the Mann–Whitney U test (Burr, 1960; Moses, 1956). Whitfield was apparently the first to discover the relationship between S, the statistic underlying Kendall's τ correlation coefficient, and U, the Mann–Whitney statistic for two independent samples. However, it was Jonckheere that was the first to make the relationship explicit (Jonckheere, 1954, p. 138). Because the Jonckheere–Terpstra test, when restricted to two independent samples, is mathematically identical in reverse application to the Wilcoxon and Mann–Whitney tests (Leach, 1979, p. 183; Randles & Wolfe, 1979, p. 396), the two-sample rank-sum test is sometimes referred to as the Kendall–Wilcoxon–Mann–Whitney–Jonckheere–Festinger test (Moses, 1956, p. 246).

Whitfield concluded his article with derivations of the variances of S for both untied and tied rankings and included a correction for continuity. For untied ranks the variance of S, as given by Kendall (1945), is $\sigma_S^2 = \{(m + n)(m + n - 1)[2(m + n) + 5]\}/18$ and the desired probability value is obtained from the asymptotically $N(0, 1)$ distribution when $\min(m, n) \to \infty$. For the example data listed in Fig. 4, the variance of S is $\sigma_S^2 = \{(2 + 4)(2 + 4 - 1)[2(2 + 4) + 5]\}/18 = 28.3333$ and $\tau = S/\sigma_S = -2/\sqrt{28.3333} = -0.3757$, with an approximate lower one-sided probability value of 0.3536.

An example will serve to illustrate the approach of Whitfield. It is common today to transform a Pearson correlation coefficient between two variables (r_{xy}) into Student's pooled t

[6]Kraft and van Eeden show how Kendall's τ can be computed as a sum of Wilcoxon W statistics (Kraft & van Eeden, 1968, pp. 180–181).

[7]Whitfield lists the date of the Kendall article as 1946, but the article was actually published in *Biometrika* in 1945.

Table 5: Fifteen paired observations with concordant/discordant (C/D) pairs and associated pair values.

Number	Pair	C/D	Value	Number	Pair	C/D	Value
1	1–2	$-,+$	-1	9	2–6	$+,+$	0
2	1–3	$-,-$	0	10	3–4	$-,-$	0
3	1–4	$-,-$	0	11	3–5	$-,-$	0
4	1–5	$-,-$	0	12	3–6	$-,+$	-1
5	1–6	$-,+$	-1	13	4–5	$-,-$	0
6	2–3	$+,-$	$+1$	14	4–6	$-,+$	-1
7	2–4	$+,-$	$+1$	15	5–6	$-,+$	-1
8	2–5	$+,-$	$+1$				

test for two independent samples and vice-versa, i.e.,

$$t = r_{xy}\sqrt{\frac{m+n-2}{1-r_{xy}^2}} \quad \text{and} \quad r_{xy} = \frac{t}{\sqrt{t^2+m+n-2}} ,$$

where m and n indicate the number of observations in Samples 1 and 2, respectively. It appears that Whitfield was the first to transform Kendall's rank-order correlation coefficient τ into Mann and Whitney's two-sample rank-sum statistic U for two independent samples. Actually, since

$$\tau = \frac{2S}{(m+n)(m+n-1)} ,$$

Whitfield established the relationship between the variable part of τ, Kendall's S, and Mann and Whitney's U. To illustrate just how Whitfield accomplished this, consider the data listed in Fig. 5. The data consist of $m = 12$ adult ages from Sample A and $n = 5$ adult ages from Sample B, with associated ranks. The sample membership of the ages/ranks is indicated by an A or a B immediately beneath the rank score. Now, arrange the two samples into a contingency table

Age:	20	20	20	20	22	23	23	24	25	25	25	25	27	29	29	35	35
Rank:	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	5	$6\frac{1}{2}$	$6\frac{1}{2}$	9	$10\frac{1}{2}$	$10\frac{1}{2}$	$10\frac{1}{2}$	$10\frac{1}{2}$	13	$14\frac{1}{2}$	$14\frac{1}{2}$	$16\frac{1}{2}$	$16\frac{1}{2}$
Sample:	A	A	A	A	B	A	A	B	A	A	A	A	B	A	A	B	B

Fig. 5: Listing of the $m + n = 17$ raw age and rank scores from Samples A and B.

with two rows and columns equal to the frequency distribution of the combined samples, as in Fig. 6. Here the first row of frequencies in Fig. 6 represents the runs in the list of ranks in Fig. 5 labeled as A, i.e., there are four values of $2\frac{1}{2}$, no value of 5, two values of $6\frac{1}{2}$, no value of 9, four values of $10\frac{1}{2}$, and so on. The second row of frequencies in Fig. 6 represents the runs in the list of ranks in Fig. 5 labeled as B, i.e., there is no value of $2\frac{1}{2}$ labeled as B, one value of 5, no value of $6\frac{1}{2}$, one value of 9, and so on. Finally, the column marginal totals are simply the sums of the two rows. This contingency arrangement permitted Whitfield to transform a problem of the difference between two independent samples into a problem of correlation between two ranked variables.

A	4	0	2	0	4	0	2	0	12
B	0	1	0	1	0	1	0	2	5
	4	1	2	1	4	1	2	2	17

Fig. 6: Contingency table of the frequency of ranks in Fig. 5.

Denote by $\mathbf{X}$ the $r \times c$ table in Fig. 6 with $r = 2$ and $c = 8$ and let x_{ij} indicate a cell frequency for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. Then, S can be expressed as the algebraic sum of all second-order determinants in $\mathbf{X}$ (Burr, 1960):

$$S = \sum_{i=1}^{r-1} \sum_{j=i+1}^{r} \sum_{k=1}^{c-1} \sum_{l=k+1}^{c} (x_{ik}x_{jl} - x_{il}x_{jk}) \ .$$

Thus, for the data listed in Fig. 6 there are $c(c-1)/2 = 8(8-1)/2 = 28$ second-order determinants:

$$S = \begin{vmatrix} 4 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} 4 & 2 \\ 0 & 0 \end{vmatrix} + \begin{vmatrix} 4 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} 4 & 4 \\ 0 & 0 \end{vmatrix} + \begin{vmatrix} 4 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} 4 & 2 \\ 0 & 0 \end{vmatrix} + \begin{vmatrix} 4 & 0 \\ 0 & 2 \end{vmatrix} + \cdots + \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} \ .$$

Therefore,

$$\begin{aligned} S = {} & (4)(1) - (0)(0) + (4)(0) - (2)(0) + (4)(1) - (0)(0) + (4)(0) - (4)(0) \\ & + (4)(1) - (0)(0) + (4)(0) - (2)(0) + (4)(2) - (0)(0) + (0)(0) - (2)(1) \\ & + (0)(1) - (0)(1) + (0)(0) - (4)(1) + (0)(1) - (0)(1) + (0)(0) - (2)(1) \\ & + (0)(2) - (0)(1) + (2)(1) - (0)(0) + (2)(0) - (4)(0) + (2)(1) - (0)(0) \\ & + (2)(0) - (2)(0) + (2)(2) - (0)(0) + (0)(0) - (4)(1) + (0)(1) - (0)(1) \\ & + (0)(0) - (2)(1) + (0)(2) - (0)(1) + (4)(1) - (0)(0) + (4)(0) - (2)(0) \\ & + (4)(2) - (0)(0) + (0)(0) - (2)(1) + (0)(2) - (0)(1) + (2)(2) - (0)(0) \end{aligned}$$

and $S = 4 + 0 + 4 + \cdots + 4 = 28$.

Alternatively, as Kendall (1948) has shown, the number of concordant pairs is given by

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} x_{ij} \left(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} x_{kl} \right)$$

and the number of discordant pairs is given by

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} x_{i,c-j+1} \left(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} x_{kl} \right) \ .$$

Thus for $\mathbf{X}$ in Fig. 6, C is calculated by proceeding from the upper-left cell with frequency $x_{11} = 4$ downward and to the right, multiplying each cell frequency by the sum of all cell frequencies below and to the right, and summing the products, i.e.,

$$\begin{aligned}C = (4)(1+0+1+0+1+0+2) + (0)(0+1+0+1+0+2) \\ + (2)(1+0+1+0+2) + (0)(0+1+0+2) \\ + (4)(1+0+2) + (0)(0+2) + (2)(2) \\ = 20+0+8+0+12+0+4 = 44 \ ,\end{aligned}$$

and D is calculated by proceeding from the upper-right cell with frequency $x_{18} = 0$ downward and to the left, multiplying each cell frequency by the sum of all cell frequencies below and to the left, and summing the products, i.e.,

$$\begin{aligned}D = (0)(0+1+0+1+0+1+0) + (2)(1+0+1+0+1+0) \\ + (0)(0+1+0+1+0) + (4)(1+0+1+0) \\ + (0)(0+1+0) + (2)(1+0) + (0)(0) \\ = 0+6+0+8+0+2+0 = 16 \ .\end{aligned}$$

Then, as defined by Kendall, $S = C - D = 44 - 16 = 28$.

To calculate Mann and Whitney's U for the data listed in Fig. 5, the number of A ranks to the left of (less than) the first B is 4; the number of A ranks to the left of the second B is 6; the number of A ranks to the left of the third B is 10; and the number of A ranks to the left of the fourth and fifth B are 12 each. Then $U = 4 + 6 + 10 + 12 + 12 = 44$. Finally, $S = 2U - mn = (2)(44) - (12)(5) = 28$. Thus Kendall's S statistic, as redefined by Whitfield, includes as special cases Yule's Q test for association in 2×2 contingency tables and the Mann–Whitney U test for larger contingency tables.

It is perhaps not surprising that Whitfield established a relationship between Kendall's S and Mann and Whitney's U as Mann published a test for trend in 1945 that was identical to Kendall's S, as Mann noted (Mann, 1945). The Mann test is known today as the Mann–Kendall test for trend where for n values in an ordered time series $x_1, \ldots, x_n$,

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(x_i - x_j) \ ,$$

where

$$\operatorname{sgn}(\cdot) = \begin{cases} +1 & \text{if } x_i - x_j > 0 \ , \\ 0 & \text{if } x_i - x_j = 0 \ , \\ -1 & \text{if } x_i - x_j < 0 \ . \end{cases}$$

7 Haldane and Smith's Rank-sum Test

John Burton Sanderson Haldane was educated at Eton and New College, University of Oxford, and was a commissioned officer during World War I. At the conclusion of the war, Haldane was awarded a fellowship at New College, University of Oxford, and then accepted a readership in biochemistry at Trinity College, University of Cambridge. In 1932 Haldane was elected a Fellow of the Royal Society and a year later, became Professor of Genetics at University College, London. In 1956, Haldane immigrated to India where he joined the Indian Statistical Institute at the invitation of P. C. Mahalanobis. In 1961 he resigned from the Indian Statistical Institute and accepted a position as Director of the Genetics and Biometry Laboratory in Orissa, India. Haldane wrote 24 books, including science fiction and stories for children, more than 400 scientific research papers, and innumerable popular articles (Mahanti, 2007). Haldane died on 1 December 1964, whereupon he donated his body to Rangaraya Medical College, Kakinada.

Cedric Austen Bardell Smith attended University College, London. In 1935, Smith received a scholarship to Trinity College, Unniversity of Cambridge, where he earned his Ph.D. in 1942. In 1946 Smith was appointed Assistant Lecturer at the Galton Laboratory, University College, London, where he met J. B. S. Haldane. In 1964 Smith accepted an appointment as the Weldon Professor of Biometry at University College, London. Cedric Smith clearly had a sense of humor and was known to occasionally sign his correspondence as "U. R. Blanche Descartes, Limit'd," which was an anagram of Cedric Austen Bardell Smith (Morton, 2002). Over his career, Smith contributed to many of the classical topics in statistical genetics, including segregation ratios in family data, kinship, population structure, assortative mating, genetic correlation, and estimation of gene frequencies (Morton, 2002).[8] Smith died on 10 January 2002, just a few weeks shy of his 85th birthday (Edwards, 2002; Morton, 2002).

In 1948 Haldane and Smith introduced an exact test for birth-order effects (Haldane & Smith, 1948). They had previously observed that in a number of hereditary diseases and abnormalities, the probability that any particular member of a sibship had a specified abnormality depended in part on his or her birth-rank (Haldane & Smith, 1948, p. 117). Specifically, they proposed to develop a quick and simple test of significance of the effect of birth-rank based on the sum of birth ranks of all affected cases in all sibships. In a classic description of a permutation test, Haldane and Smith noted that if in each sibship the numbers of normal and affected siblings were held constant, then if birth-rank had no effect, every possible order of normal and affected siblings would be equally probable. Accordingly, the sum of birth-ranks for affected siblings would have a definite distribution, free from unknown parameters, providing "a 'conditional' and 'exact' test for effect of birth-rank" (Haldane & Smith, 1948, p. 117). Finally, they observed that this distribution would be very nearly normal in any practically occurring case with a mean and variance that were easily calculable.

Consider a single sibship of k births, h of which are affected.[9] Let the birth-ranks of the affected siblings be denoted by $a_1,\ a_2,\ldots,a_h$ and their sum by $A = \sum_{r=1}^{h} a_r$. Then, there are

$$\binom{k}{h} = \frac{k!}{h!(k-h)!}$$

equally-likely ways of distributing the h affected siblings. Of these, the number of ways of distributing them, $P_{h,\,k}(A)$, so that their birth-ranks sum to A is equal to the number of partitions

[8]Cedric Smith, Roland Brooks, Arthur Stone, and William Tutte met at Trinity College, University of Cambridge, and were known as the Trinity Four. Together they published mathematical papers under the pseudonym Blanche Descartes, much in the tradition of the putative Peter Ørno, John Rainwater, and Nicolas Bourbaki.

[9]Haldane and Smith used k for the number of births and h for the number of affected births, instead of the more conventional n and r, respectively.

of A into h unequal parts, $a_1,\ a_2, \ldots, a_h$, no part being greater than k. Given this, the probability $p_{h,\,k}(A)$ of obtaining a sum A is given by

$$p_{h,\,k}(A) = P_{h,\,k}(A) \Bigg/ \binom{k}{h} \ . \tag{7.1}$$

Dividing these partitions into two classes according to whether the greatest part is or is not k, yields

$$P_{h,\,k}(A) = P_{h,\,k-1}(A) + P_{h-1,\,k-1}(A-k) \ . \tag{7.2}$$

Haldane and Smith observed that from the relation described in Eq. 7.2 they could readily calculate $P_{h,\,k}(A)$ for small samples of h and k.

Since $(k+1-a_1),\ (k+1-a_2), \ldots, (k+1-a_h)$ must be a set of h integers, all different, not greater than k, and summing to $h(k+1)-A$, they showed that

$$P_{h,\,k}(A) = P_{h,\,k}[h(k+1)-A] \ . \tag{7.3}$$

Haldane and Smith went on to note that, similar to the affected siblings, in any sibship the unaffected siblings would all have different birth-ranks, none exceeding k, but summing to $\frac{k}{2}(k+1)-A$. Thus,

$$P_{h,\,k}(A) = P_{k-h,\,k}\left[\tfrac{k}{2}(k+1)-A\right] \ . \tag{7.4}$$

An example will serve to illustrate the recursive process of Haldane and Smith.[10] Consider a sibship of $k = 6$ siblings with $h = 2$ of the siblings classified as affected (a) and $k-h = 6-2 = 4$ of the siblings classified as normal (n), with birth-order indicated by subscripts: $n_1,\ a_2,\ n_3,\ n_4,\ a_5,\ n_6$. Thus, the affected siblings are the second and fifth born out of six siblings and yield a sum of $A = a_2 + a_5 = 2+5 = 7$. Table 6 lists the partitions and associated frequency distributions for $h = 2$ and $k = 6$ in the first set of columns, $h = 2$ and $k-1 = 6-1 = 5$ in the second set of columns, $h-1 = 2-1 = 1$ and $k-1 = 6-1 = 5$ in the third set of columns, and $k-h = 6-2 = 4$ and $k = 6$ in the fourth set of columns. It can be seen in Table 6 that $P_{h,\,k}(A) = P_{2,\,6}(7) = 3$ since there are three ways of placing an affected sibling yielding a sum of $A = 7$, i.e, $\{1,\ 6\}$, $\{2,\ 5\}$, and $\{3,\ 4\}$. As there are a total of

$$\binom{k}{h} = \frac{k!}{h!(k-h)!} = \frac{6!}{2!(6-2)!} = 15$$

equally probable ways of placing the $h = 2$ affected siblings, the probability of obtaining a sum of $A = 7$ as given in Eq. 7.1 is

$$p_{2,\,6}(7) = P_{2,\,6}(7)/15 = 3/15 = 0.20 \ .$$

Dividing the partitions into two classes as in Eq. 7.2 yields

$$\begin{aligned} P_{2,\,6}(7) &= P_{2,\,6-1}(7) + P_{2-1,\,6-1}(7-6) \ , \\ 3 &= P_{2,\,5}(7) + P_{1,\,5}(1) \ , \\ 3 &= 2+1 \ , \end{aligned}$$

[10]It should be noted that while the decomposition in Eq. 7.4 is different than that employed by Mann and Whitney (1947) in Eq. 5.1, it is similar to the decomposition used by Festinger (1946).

Table 6: Partitions (P), sums (A), and frequencies (f) for $P_{h,k}(A) = P_{2,6}(7)$, $P_{h,k-1}(A) = P_{2,5}(7)$, $P_{h-1,k-1}(A-k) = P_{1,5}(1)$, and $P_{k-h,k}[\frac{k}{2}(k+1) - A] = P_{4,6}(14)$.

$P_{2,6}(7)$			$P_{2,5}(7)$			$P_{1,5}(1)$			$P_{4,6}(14)$		
P	A	f	P	A	f	P	A	f	P	A	f
1, 2	3	1	1, 2	3	1	1	1	1	1, 2, 3, 4	10	1
1, 3	4	1	1, 3	4	1	2	2	1	1, 2, 3, 5	11	1
1, 4	5	2	1, 4	5	2	3	3	1	1, 2, 3, 6	12	2
1, 5	6	2	1, 5	6	2	4	4	1	1, 2, 4, 5	13	2
1, 6	7	3	2, 3	7	2	5	5	1	1, 2, 4, 6	14	3
2, 3	8	2	2, 4	8	1				1, 2, 5, 6	15	2
2, 4	9	2	2, 5	9	1				1, 3, 4, 5	16	2
2, 5	10	1	3, 4						1, 3, 4, 6	17	1
2, 6	11	1	3, 5						1, 3, 5, 6	18	1
3, 4			4, 5						1, 4, 5, 6		
3, 5									2, 3, 4, 5		
3, 6									2, 3, 4, 6		
4, 5									2, 3, 5, 6		
4, 6									2, 4, 5, 6		
5, 6									3, 4, 5, 6		

as illustrated in Table 6, where $P_{2,6}(7)$ in the first set of columns is associated with a frequency of 3, $P_{2,5}(7)$ in the second set of columns is associated with a frequency of 2, and $P_{1,5}(7-6) = P_{1,5}(1)$ in the third set of columns is associated with a frequency of 1; thus, $3 + 2 + 1$. Note that once again, the decomposition observed in the discussion of Festinger and the two-sample rank sum test appears wherein

$$\binom{k}{h} = \binom{k-1}{h} + \binom{k-1}{h-1} ,$$

$$\binom{6}{2} = \binom{6-1}{2} + \binom{6-1}{2-1} ,$$

$$\binom{6}{2} = \binom{5}{2} + \binom{5}{1} ,$$

$$15 = 10 + 5 .$$

This decomposition can be observed in Table 6, where the column of frequencies for $P_{2,6}(A)$ in the first set of columns sums to 15, the column of frequencies for $P_{2,5}(A)$ in the second set of columns sums to 10, the column of frequencies for $P_{1,5}(A-k)$ in the third set of columns sums to 5, and $15 = 10 + 5$.

From Eqs. 7.3 and 7.4, Haldane and Smith were able to construct a table of values of $P_{h,k}(A)$ and $\binom{k}{h}$, giving the exact distribution for all values of k up to and including 12, noting that values not explicitly given in the table could readily be derived by the use of Eqs. 7.3

and 7.4. Additionally, Haldane and Smith investigated the approximate distribution of A. They found it more efficient to test $6A$ instead of A and showed that the theoretical mean of $6A$ was $3h(k+1)$ and the theoretical variance was $3h(k+1)(k-h)$, and thus provided a table of means and variances for $h = 1, \ldots, 18$ and $k = 2, \ldots, 20$. They observed that since A is made up of a number of independent components, the distribution of A would be approximately normal and, therefore, if an observed value of A exceeded the mean by more than twice the standard deviation, siblings born later were most likely to be affected, but if the observed value of A fell short of the mean by the same amount, siblings born earlier were most likely to be affected (Haldane & Smith, 1948, p. 121). They concluded the paper with an example analysis based on data from Munro on phenylketonuria from forty-seven British families (Munro, 1947).

8 van der Reyden's Rank-sum Test

Little is known about D. van der Reyden other than that he was a statistician for the Tobacco Research Board of Southern Rhodesia.[11] In 1952 van der Reyden independently developed a two-sample rank-sum test equivalent to the tests of Wilcoxon (1945), Festinger (1946), Mann and Whitney (1947), Whitfield (1947), and Haldane and Smith (1948), although none of these is referenced; in fact, the article by van der Reyden contains no references whatsoever. The stated purpose of the proposed test was to provide a simple exact test of significance using sums of ranks in order to avoid computing sums of squares (van der Reyden, 1952, p. 96). In a novel approach, van der Reyden utilized a tabular format involving rotations of triangular matrices to generate permutation distribution frequencies and published tables for tests of significance at the 0.05, 0.02, and 0.01 levels (van der Reyden, 1952).

Table 7 illustrates the van der Reyden tabular procedure with values of $n = 1, 2, 3$, $m = 1, \ldots, 6$, and sums of frequencies from $T = 1$ to $T = 15$. Looking first at the column headed $n = 1$ in Table 7, note that when $m = 1$ and $n = 1$, $T = 1$; when $m = 2$ and $n = 1$, $T = 1$ or 2; when $m = 3$ and $n = 1$, $T = 1$, 2, or 3; and when $m = 4$ and $n = 1$, $T = 1$, 2, 3, or 4. Simply put, taking all samples of one item from m items, all values of T will have a frequency of 1. In this case, each T has a frequency of 1 and each frequency sums to $\binom{m}{n}$, e.g., for $m = 4$ and $n = 1$ the frequency distribution is $\{1, 1, 1, 1\}$ with a sum of 4, which is $\binom{4}{1} = 4$. To obtain the frequencies for samples of $n = 2$ items, rotate all frequencies for $n = 1$ clockwise through $45°$, shifting the whole distribution downward to

$$T = \binom{n+1}{2} = \frac{n(n+1)}{2} \; .$$

Thus in Table 7, the frequencies obtained for $n = 1$ are transposed with the first row now constituting the fourth column, the second row constituting the third column, and so on. Then this transposed matrix is shifted downward so it begins at $T = n(n+1)/2 = 2(2+1)/2 = 3$. Finally, the frequencies are added together horizontally in the same manner as Festinger (1946), as follows.

Consider the frequency distributions listed under $n = 2$ in Table 7. There are two sets of frequency distributions under $n = 2$, one on the left and one on the right, both labeled

[11] Southern Rhodesia was shortened to Rhodesia in 1965 and renamed the Republic of Zimbabwe in 1980.

Table 7: Generation of frequency arrays for $n = 1$, $n = 2$, and $n = 3$ as described by van der Reyden.

	$n=1$				$n=2$								$n=3$							
T/m	1	2	3	4	2	3	4	5	2	3	4	5	3	4	5	6	3	4	5	6
1	1	1	1	1																
2		1	1	1																
3			1	1	1				1	1	1	1								
4				1		1				1	1	1								
5						1	1			1	2	2								
6							1	1			1	2	1				1	1	1	1
7							1	1			1	2		1				1	1	1
8								1				1		1	1			1	2	2
9								1				1		1	1	1		1	2	3
10															2	1			2	3
11															1	2			1	3
12															1	2			1	3
13																2				2
14																1				1
15																1				1

$m = 2, 3, 4, 5$. So, for example, to create the frequency distribution listed under $n = 2$, $m = 3$ on the right, add together the frequency distribution listed under $n = 2$, $m = 2$ on the right and the frequency distribution under $n = 2$, $m = 3$ on the left. To create the frequency distribution listed under $n = 2$, $m = 4$ on the right, add together the frequency distribution listed under $n = 2$, $m = 3$ on the right and the frequency distribution under $n = 2$, $m = 4$ on the left. To create the frequency distribution listed under $n = 2$, $m = 5$ on the right, add together the frequency distribution listed under $n = 2$, $m = 4$ on the right and the frequency distribution under $n = 2$, $m = 5$ on the left. The process continues in this manner, recursively generating the required frequency distributions.

For a final example, consider the frequency distributions listed under $n = 3$ in Table 7. Again there are two sets of frequency distributions, one on the left and one on the right. The distribution on the left is created by rotating the distribution created under $n = 2$ on the right, and shifting it downward so it begins at $T = n(n + 1)/2 = 3(3 + 1)/2 = 6$. To create the frequency distribution listed under $n = 3$, $m = 6$ on the right, add together the frequency distribution listed under $n = 3$, $m = 5$ on the right and the frequency distribution under $n = 3$, $m = 6$ on the left. The frequency distributions of sums in Table 7 can be compared with the frequency distributions of sums in Table 3 that were generated with Festinger's method. In this recursive manner van der Reyden created tables for T from $n = 2, \ldots, 12$ and $m = 10, \ldots, 30$ for the α = 0.05, 0.02, and 0.01 levels of significance.

9 Further Developments

It would be remiss not to mention a later contribution by S. Siegel and J. W. Tukey. In 1960 Siegel and Tukey developed a non-parametric two-sample test based on differences in variability between the two unpaired samples, rather than the more conventional tests for differences in location (Siegel & Tukey, 1960). The Siegel–Tukey test was designed to replace parametric F tests for differences in variances that depended heavily on normality, such as Bartlett's F and Hartley's $F_{\max}$ tests for homogeneity of variance (Bartlett, 1937; Hartley, 1950). Within this article Siegel and Tukey provided tables of one- and two-sided critical values based on exact probabilities for a number of levels of significance.

Let the two sample sizes be denoted by n and m with $n \leq m$ and assign ranks to the $n + m$ ordered observations with low ranks assigned to extreme observations and high ranks assigned to central observations. Since the sum of the ranks is fixed, Siegel and Tukey chose to work with the sum of ranks for the smaller of the two samples, represented by R_n. They also provided a table with one- and two-sided critical values of R_n for $n \leq m \leq 20$ for various levels of α.

Siegel and Tukey noted that their choice of ranking procedure, with low ranks assigned to extreme observations and high ranks assigned to central observations, allowed the use of the same tables as were used for the Wilcoxon two-sample rank-sum test for location. Thus, as they explained, their new test "[might] be considered a Wilcoxon test for spread in unpaired samples" (Siegel & Tukey, 1960, p. 432). Alternatively, as they noted, the Siegel–Tukey tables were equally applicable to the Wilcoxon (1945), Festinger (1946), Mann–Whitney (1947), and White (1952) rank-sum procedures for relative location of two unpaired samples, and were appropriate linear transformations of the tabled values presented by Auble (1953).

In addition, a number of tables of exact probability values and/or tests of significance were explicitly constructed for the two-sample rank-sum test in succeeding years, all based on either the Wilcoxon or the Mann–Whitney procedure. By 1947 Wilcoxon had already published extended probability tables for the Wilcoxon W statistic (Wilcoxon, 1947). In 1952 C. White utilized the "elementary methods" of Wilcoxon to develop tables for the Wilcoxon W statistic when the numbers of items in the two independent samples were not necessarily equal (White, 1952). In 1953 D. Auble published extended tables for the Mann–Whitney U statistic (Auble, 1953) and in 1955 E. Fix and J. L. Hodges published extended tables for the Wilcoxon W statistic (Fix & Hodges, 1955). In 1963 J. E. Jacobson published extensive tables for the Wilcoxon W statistic (Jacobson, 1963) and in 1964 R. C. Milton published an extended table for the Mann–Whitney U statistic (Milton, 1964).

There were a number of errors in several of the published rank-sum tables that were corrected in later articles; see especially the 1963 article by L. R. Verdooren (Verdooren, 1963) that contained corrections for the tables published by White (White, 1952) and Auble (Auble, 1953), and an erratum to the 1952 article by Kruskal and Wallis (Kruskal & Wallis, 1952) that corrected errors in the tables published by White (White, 1952) and van der Reyden (van der Reyden, 1952). By the mid-1960s tables of exact probability values were largely supplanted by efficient computer algorithms.

References

Auble, D. (1953) Extended tables for the Mann–Whitney statistic, *Bulletin of the Institute of Educational Research*, **1**, 1–39.

Bartlett, M. S. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A (Mathematical and Physical Sciences)*, **160**, 268–282.

Bradley, R. A. (1966) Frank Wilcoxon, *Biometrics*, **22**, 192–194.

__________ (1997) Frank Wilcoxon. In N. L. Johnson, and S. Kotz (eds.) *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*, Wiley: New York, pp. 339–341.

__________ and M. Hollander (2001) Wilcoxon, Frank. In C. C. Heyde and E. Seneta (eds.) *Statisticians of the Centuries*, Springer–Verlag: New York, pp. 420–424.

Burr, E. J. (1960) The distribution of Kendall's score S for a pair of tied rankings. *Biometrika*, **47**, 151–171.

Deuchler, G. (1914) Über die methoden der korrelationsrechnung in der pädogogik und psychologie, *Zeitschrift für Pädagogische Psychologie und Experimentelle Pädagogik*, **15**, 114–131, 145–159, and 229–242.

Edwards, A. W. F. (2002) Professor C. A. B. Smith, 1917 – 2002, *Journal of the Royal Statistical Society, Series D (The Statistician)*, **51**, 404–405.

Festinger, L. (1946) The significance of differences between means without reference to the frequency distribution function, *Psychometrika*, **11**, 97–105.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*, Oliver and Boyd: Edinburgh.

Fix, E. and J. L. Hodges (1955) Significance probabilities of the Wilcoxon test, *The Annals of Mathematical Statistics*, **26**, 301–312.

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association*, **32**, 675–701.

Gini, C. (1916/1959) Il concetto di "Transvariazione" e le sue prime applicazioni, *Giornale degli economisti e Rivista di Statistica*, **53**, 13–43. Reproduced in Gini, C. (1959) *Transvarizione. Memorie de Metodologia Statistica: Volume Secundo: a cura di Giuseppe Ottaviani*, Libreria Goliardia: Rome.

Haldane, J. B. S. and C. A. B. Smith (1948) A simple exact test for birth-order effect, *Annals of Eugenics*, **14**, 117–124.

Hartley, H. O. (1950) The use of range in analysis of variance, *Biometrika*, **37**, 271–280.

Hollander, M. (2000) A conversation with Ralph A. Bradley, *Statistical Science*, **16**, 75–100.

Jacobson, J. E. (1963) The Wilcoxon two-sample statistic: Tables and bibliography, *Journal of the American Statistical Association*, **58**, 1086–1103.

Jonckheere, A. R. (1954) A distribution-free k-sample test against ordered alternatives. *Biometrika*, **41**, 133–145.

Kendall, M. G. (1938) A New Measure of Rank Correlation, *Biometrika*, **30**, 81–93.

____________ (1945) The treatment of ties in ranking problems, *Biometrika*, **33**, 239–251.

____________ (1948) *Rank Correlation Methods*, Charles Griffin: London.

Kruskal, W. H. (1957) Historical notes on the Wilcoxon unpaired two-sample test, *Journal of the American Statistical Association*, **52**, 356–360.

____________ and W. A. Wallis (1952) Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, **47**, 583–621. Erratum: Journal of the American Statistical Association, **48**, 907–911 (1953).

Leach, C. (1979) *Introduction to Statistics: A Nonparametric Approach for the Social Sciences*, John Wiley & Sons: New York.

Lipmann, O. (1908) Eine methode zur vergleichung zwei kollektivgegenständen, *Zeitschrift fur Psychologie*, **48**, 421–431.

MacMahon, P. A. (1916) *Combinatory Analysis*, Vol. II, Cambridge University Press: Cambridge.

Mahanti, S. (2007) John Burdon Sanderson Haldane: The ideal of a polymath, *Vigyan Prasar Science Portal*. Accessed on 20 January 2012 at `http://www.vigyanprasar.gov.in/scientists/JBSHaldane.htm`.

Mann, H. B. (1945) Nonparametric test against trend, *Econometrica*, **13**, 245–259.

__________ and D. R. Whitney (1947) On a test of whether one of two random variables is stochastically larger than the other, *The Annals of Mathematical Statistics*, **18**, 50–60.

Milton, R. C. (1964) An extended table of critical values for the Mann–Whitney (Wilcoxon) two-sample dtatistic, *Journal of the American Statistical Association*, **59**, 925–934.

Morton, N. (2002) Cedric Smith (1917 – 2002), *International Statistical Institute Newsletter*, **26**, 9–10.

Moscovici, S. (1989) Obituary: Leon Festinger, *European Journal of Social Psychology*, **19**, 263–269.

Moses, L. E. (1956) Statistical theory and research design. *Annual Review of Psychology*, **7**, 233–258.

Munro, T. A. (1947) Phenylketonuria: Data on forty-seven British families, *Annals of Human Genetics*, **14**, 60–88.

Olson, J. (c. 2000) Henry Berthold Mann, *Department of Mathematics, The Ohio State University*. Accessed on 20 January 2012 at `http://www.math.osu.edu/history/biographies/mann`.

Ottaviani, G. S. (1939) Probabilitá che una prova su due variabili casuali X e Y verifichi la disuguaglianza $X < Y$ e sul Corrispondente Scarto Quadratico Medio, *Giornale dell'Istituto Italiano degli Attuari*, **10**, 186–192.

Pascal, B. (1665/1959) Traité du triangle arithmétique (Treatise on the arithmetical triangle). In D. E. Smith (ed.) *A Source Book in Mathematics*, Vol. 1, Dover, New York, pp. 67–79. Translated by A. Savitsky.

Potvin, C. and D. A. Roff (1993) Distribution-free and robust statistical methods: Viable alternatives to parametric statistics, *Ecology*, **74**, 1615–1628.

Randles, R. H. and D. A. Wolfe (1979) *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons: New York.

Siegel, S. and J. W. Tukey (1960) A nonparametric sum of ranks procedure for relative spread in unpaired samples, *Journal of the American Statistical Association*, **55**, 429–445.

Stigler, S. M. (1999) *Statistics on the Table: The History of Statistical Concepts and Methods*, Harvard University Press: Cambridge, MA.

van der Reyden, D. (1952) A simple statistical significance test, *The Rhodesia Agricultural Journal*, **49**, 96–104.

Vankeerberghen, P., C. Vandenbosh, J. Smeyers-Verbeke, and D. L. Massart (1991) Some robust statistical procedures applied to the analysis of chemical data, *Chemometrics and Intelligent Laboratory Systems*, **12**, 3–13.

Verdooren, L. R. (1963) Extended tables of critical values for Wilcoxon's test statistic, *Biometrika*, **50**, 177–186.

White, C. (1952) The use of ranks in a test of significance for comparing two treatments, *Biometrics*, **8**, 33–41.

Whitfield, J. W. (1947) Rank correlation between two variables, one of which is ranked, the other dichotomous, *Biometrika*, **34**, 292–296.

Whitworth, W. A. (1942) *Choice and Chance*, New York: G. E. Stechert.

Wilcoxon, F. (1945) Individual comparisons by ranking methods, *Biometrics Bulletin*, **1**, 80–83.

__________ (1947) Probability tables for individual comparisons by ranking methods, *Biometrics*, **3**, 119–122.

Willke, T. (2008) In Memoriam – Ransom Whitney, *The Ohio State University, Department of Statistics News*, **16**. Accessed on 17 January 2012 at `http://www.stat.osu.edu/sites/default/files/news/statnews2008.pdf`.