RANK METHODS FOR COMBINATION OF INDEPENDENT EXPERIMENTS IN ANALYSIS OF VARIANCE[1]

By J. L. Hodges, Jr. and E. L. Lehmann

University of California, Berkeley

1. Introduction and summary. It is now coming to be generally agreed that in testing for shift in the two-sample problem, certain tests based on ranks have considerable advantage over the classical t-test. From the beginning, rank tests were recognized to have one important advantage: their significance levels are exact under the sole assumption that the samples are randomly drawn (or that the assignment of treatments to subjects is performed at random), whereas the t-test in effect is exact only when we are dealing with random samples from normal distributions. On the other hand, it was felt that this advantage had to be balanced against the various optimum properties possessed by the t-test under the assumption of normality. It is now being recognized that these optimum properties are somewhat illusory and that, under realistic assumptions about extreme observations or gross errors, the t-test in practice may well be less efficient than such rank tests as the Wilcoxon or normal scores test [6], [7].

Rank tests were naturally developed first for the simple two-sample problem,[2] but in practice experiments for the evaluation of population or treatment differences are seldom of this form. To secure the advantages of increased homogeneity and the resultant increased precision, it is more customary to stratify the populations or to divide the experimental subjects into blocks, using a (generalized) randomized block design. In other experiments where additivity of certain effects is assumed, a design of the Latin square type may be appropriate.

It is the purpose of the present paper to provide a method for constructing rank tests for such designs. The basic idea of the method is described in Section 3. The main body of the paper is concerned with the application of the method to the comparison of two treatments using a Wilcoxon-type test statistic. The exact distribution of this statistic is discussed in Section 4, and its asymptotic distribution in Sections 5 and 6. Some remarks concerning the efficiency of the test are given in Section 7. Finally, Section 8 illustrates the application of the method to the comparsion of more than two treatments, for such designs as incomplete blocks and Latin squares.

2. Tests based on independent rankings. A simple method for dealing with at least some cases of the problem described in the introduction is based on separate rankings of the observations in each block.

Received January 30, 1960.

[1] This paper was prepared with the partial support of the Office of Naval Research (nonr-222-43). This paper in whole or in part may be reproduced for any purpose of the United States Government.

[2] And for the case of paired comparisons, which also turns out to lend itself particularly well to a rank analysis.

J. Rojo (ed.), *Selected Works of E. L. Lehmann*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1412-4_35, © Springer Science+Business Media, LLC 2012

To fix ideas, let us consider the generalized randomized block design for comparing two treatments. Suppose that N subjects are available for the experiment. They are divided into n blocks, each block consisting of subjects thought to be homogeneous with regard to the response to be observed. The N_i subjects in the ith block are then randomly divided into groups of sizes s_i and t_i ($s_i + t_i = N_i$), and the groups are given treatments A and B respectively. We want a test of the hypothesis that the treatments do not differ, which will be sensitive to the shift alternatives that the response to treatment B tends consistently to be higher than the response to A.

A rank test for this problem was proposed by Wilcoxon [10]. The observations in each block are ranked separately, a Wilcoxon statistic (for example, the sum of the B-ranks) $W^{(i)}$ is formed for the ith block, and the test statistic is the sum of these W's. In a recent investigation [3], van Elteren has considered the more general statistic $\sum_{i=1}^{n} c_i W^{(i)}$ and has shown that in a certain sense the optimum constants are

$$c_i = 1/(s_i + t_i + 1). \tag{2.1}$$

What is the efficiency of this test relative to the appropriate t- or normal test, which is based on the statistic

$$\sum \frac{s_i t_i}{s_i + t_i} (Y_{i\cdot} - X_{i\cdot}) \Big/ \sum \frac{s_i t_i}{s_i + t_i} \tag{2.2}$$

where $X_{i\cdot}$ and $Y_{i\cdot}$ denote the average of the A- and B-responses in the ith block? To answer this, suppose that the Wilcoxon experiment consists of r' replications of some given set of blocks (s_i, t_i), and that the t-experiment consists of $r'' = g(r')$ replications of the same set where r'' will be determined below. Suppose that the X's and Y's in the ith block are normally distributed with means ξ_i and $\xi_i + \Delta$ respectively and with common variance 1. For fixed values of α and β, let $\Delta = \Delta(r')$ be determined so that the power of the Wilcoxon-van Elteren level α test tends to β as $r' \to \infty$. Let $r'' = g(r')$ be determined, so that the power of the level α t-test (based on r'' replications) against the same sequence of alternatives $\Delta(r')$ also tends to β. We then say that the relative asymptotic efficiency of the Wilcoxon-van Elteren test to the t-test is e if

$$\lim_{r' \to \infty} (r'/r'') = e,$$

independent of α and β.

THEOREM 1. *For block sizes (s_i, t_i) the relative asymptotic efficiency of the Wilcoxon-van Elteren to the t-test based on the statistic* (2.2) *is*

$$e = \frac{3}{\pi} \sum \frac{s_i t_i}{s_i + t_i + 1} \Big/ \sum \frac{s_i t_i}{s_i + t_i}. \tag{2.3}$$

PROOF. Since the method of proof is well known (see for example [7]), it will be enough here to sketch the proof. The statistic (2.2) (and therefore asymptotically as $\sum s_i$ and $\sum t_i \to \infty$ also the corresponding t-statistic) is normally distributed with mean Δ and variance $[\sum s_i t_i/(s_i + t_i)]^{-1}$.

Analogously, the Wilcoxon statistic in the ith block, for small Δ has approximately mean $s_it_i(\frac{1}{2} + \frac{1}{2}\Delta/\sqrt{\pi})$ and null variance $s_it_i(s_i + t_i + 1)/12$ and the linear combination of these statistics with weights (2.1) therefore has approximately mean $(\frac{1}{2} + \frac{1}{2}\Delta/\sqrt{\pi}) \sum s_it_i/(s_i + t_i + 1)$ and null variance

$$\sum s_it_i(s_i + t_i + 1)/12.$$

If the Wilcoxon statistics are based on (s_i', t_i') and the t-test on (s_i'', t_i'') observations, and if $\sum s_i'$, $\sum t_i'$ are sufficiently large for the normal approximation to be close, the power of the two tests for small Δ will be approximately the same if

$$\Delta^2 \sum \frac{s_i'' t_i''}{s_i'' + t_i''} = \frac{\Delta^2}{4\pi} \frac{\left(\sum \frac{s_i' t_i'}{s_i' + t_i' + 1}\right)^2}{\frac{1}{12} \sum \frac{s_i' t_i'}{s_i' + t_i' + 1}},$$

that is, if

$$\sum \frac{s_i'' t_i''}{s_i'' + t_i''} = \frac{3}{\pi} \sum \frac{s_i' t_i'}{s_i' + t_i' + 1}. \tag{2.4}$$

Suppose now that the Wilcoxon experiment consists of r' replications of some given set of blocks (s_i, t_i); and that the t-experiment consists of r'' replications of the given set. Then (2.4) becomes

$$r'' \sum \frac{s_i t_i}{s_i + t_i} = r' \frac{3}{\pi} \sum \frac{s_i t_i}{s_i + t_i + 1}$$

and this completes the proof.

We note two interesting special cases of (2.3). If the block size $N_i = s_i + t_i$ is constant, $N_i = k$ say, then (2.3) reduces to

$$e_k = \frac{3}{\pi} \frac{k}{k + 1} \tag{2.5}$$

regardless of how the numbers s_i, t_i are chosen in each block. Table 2.1 gives this efficiency for a number of different values of k.

Another case of interest is that of a single block. The test then is the ordinary Wilcoxon test, and as $s + t \to \infty$, the efficiency $(3/\pi)(s + t)/(s + t + 1)$ tends to the value $3/\pi$, ordinarily given as the asymptotic relative efficiency of the Wilcoxon to the t-test in this case.

For the comparison of more than two, say $c + 1$, treatments in a randomized block design ($c + 1$ subjects in each block to which the $c + 1$ treatments are assigned at random), an analogous rank test was proposed by Friedman [5],[3] and his test was generalized to balanced incomplete randomized blocks by Durbin [2] and to general blocks by Benard and van Elteren [1]. In these tests,

[3] For other rank procedures for this problem see Walsh [9].

TABLE 2.1

k	2	3	4	5	10	15	∞
e_k	$\frac{2}{\pi} = .637$	.716	.764	.796	.868	.895	$\frac{3}{\pi} = .955$

the ranking is done separately within each block, and the test statistic is a suitable quadratic form in the sum of the ranks for the various treatments. When the block size is k, it was shown by van Elteren and Noether [4] that the asymptotic relative efficiencies of Friedman's and Durbin's tests relative to the appropriate F-tests is also given by (2.5).

For the case of two treatments and blocks of size $k = 2$, the efficiency $2/\pi$ is not surprising since the Wilcoxon sum and Friedman's test then reduce to the one- or two-sided sign test for matched pairs. It is seen from the table of e_k that the efficiency remains unpleasantly low as long as the blocks are small. This is unfortunate since it is often desirable to use rather small blocks either because the natural blocks are small (for example, litters) or because small blocks are required to achieve within-block homogeneity. In such cases, tests based on independent rankings leave much to be desired.

There are of course situations in which separate rankings are all that can be obtained. An example is the case in which the blocks correspond to different observers, each of whom makes a comparison of the different treatments assigned at random to different sets of subjects. If this comparison can be made only in the form of a ranking, the basic data are just the separate rankings and the over-all evaluation must be based on these data. However, if instead each observer assigns scores to the different treatments, even if the method of scoring is not the same for each observer, the method to be outlined in the next section can be applied.

3. Ranking after alignment. While in the case of a matched pairs experiment ($s_i = t_i = 1$), the efficiency of the Wilcoxon-sum test is only $2/\pi$, there does of course in this case exist a rank test, Wilcoxon's one-sample test, whose efficiency is $3/\pi$ in normal populations. In this test the absolute differences in response are ranked for the n pairs, and with each rank is associated the sign of the corresponding response difference. The test statistic is the sum of the ranks corresponding to the positive (or negative) differences.

It seems natural to ask why this test has efficiency $3/\pi$ as compared with the efficiency $2/\pi$ for the Wilcoxon-sum test. The reason appears to be essentially that the former test pays attention to certain interblock comparisons which are entirely ignored by the latter test. The main objective of the present paper is to find rank procedures for more general designs that will preserve the interblock comparisons in the hope that this will lead to higher efficiency.

We can best introduce the method by continued consideration of the com-

TABLE 3.1

Block	Treatment A	Treatment B	Mean
1	98, 169	28, 113	102
2	259	168, 128	185
3	81, 120	24, 102, 8	67

parison of two treatments. Suppose that treatments A and B are compared by random assignment to the subjects in three blocks with $s_1 = t_1 = 2$; $s_2 = 1$, $t_2 = 2$; $s_3 = 2$, $t_3 = 3$, and that the results are observed as given in Table 3.1. From each observation we subtract the block mean, arrange the residuals in order of increasing size, and rank them:

Residual	−74	−59	−57	−43	−17	−4	11	14	35	53	67	74
Rank	1	2	3	4	5	6	7	8	9	10	11	12

The ranks corresponding to treatment A are underlined; their sum

$$6 + 8 + 10 + 11 + 12 = 47$$

is the value of a test statistic which will be denoted by W.

Let us first observe that W provides an exact test of the null hypothesis that the treatments do not differ. Under that hypothesis, the labels A and B are meaningless and could as well be assigned at the end of the experiment, after the data are recorded. Each block has certain ranks, as tabulated below.

Block	Ranks
1	1, 6, 7, 11
2	3, 5, 12
3	2, 4, 8, 9, 10

The random assignment of two A-labels to the four ranks of Block 1 produces two A-ranks, whose sum will be denoted by W_1. W_1 is equally likely to have any of the values $1 + 6 = 7$, $1 + 7 = 8$, $1 + 11 = 12$, $6 + 7 = 13$, $6 + 11 = 17$ or $7 + 11 = 18$. Similarly W_2, the rank labeled A in Block 2, takes on the values 3, 5, 12 with probability $\frac{1}{3}$ each, while W_3 is equally likely to have any of the 10 values $2 + 4 = 6$, 10, 11, 12, 12, 13, 14, 17, 18, 19. The statistic $W = W_1 + W_2 + W_3$ is the sum of three independent integer-valued random variables. The maximum value of W is 49, and one sees by inspection that the only value of W_1, W_2, W_3 for which $W \geqq 47$ are $W_1 = 17$ or 18, $W_2 = 12$, $W_3 = 17$, 18 or 19. Thus

$$P(W \geqq 47) = \frac{2 \cdot 3}{\binom{4}{2}\binom{3}{1}\binom{5}{2}} = \frac{6}{180} = \frac{1}{30}.$$

We can now state the idea of the procedure. The first step is to bring the observations in the various blocks into alignment with one another. In the ex-

ample above this was done by subtracting from each observation the mean observation in its block, but in some cases other methods for alignment might be better, such as subtracting a trimmed or Winsorized mean.[4] The important point is that the treatments must be ignored when the alignment is made.

Once the observations are aligned they are pooled and ranked without regard to their blocks. Then the ranks are labeled according to the treatment given to the corresponding observation. Under the null hypothesis of no treatment effect, the assignment of labels to the ranks pertaining to each block is done at random, and may be thought of as having been done after the ranks for each block are determined. The partition of the ranks in each block into label-groups is independent, and has in each block a known distribution that depends only on the design employed at the beginning of the experiment in assigning treatments to subjects.

Finally, we may compute from the labeled ranks any rank statistic appropriate for the alternatives against which it is desired to test the null hypothesis. In the example we used the Wilcoxon statistic and computed its one-sided significance probability, but one may in other cases want a two-sided test, or use the normal scores test statistic, or indeed whatever statistic seems appropriate. In any case, the statistic will have an exact null distribution whose computation depends on the known distribution of the labels among the block ranks.

The procedure is thus very flexible. We have already mentioned that the method of alignment and the choice of test statistic may be adjusted to the problem. Still another possibility is that of first transforming the data, where a different transformation may be used in each block. None of these devices affects the exactness of the null distribution. They do however influence the power of the test, and the choices should be made with the view of providing large power against the alternatives of interest.

The null distributions of all the tests suggested above are based on the independent random assignment of treatments in each block. One can of course also apply the tests to a different experimental situation, in which the subjects in the different blocks are randomly sampled from different populations, but in this case the null distributions discussed above are conditional distributions, given the responses of the sampled subjects.

4. Rank sum analysis of two treatments in a block design. The computation of significance probability in the example considered above was carried out by inspection. The various equally likely samples were examined to determine the number of them at least as significant as the one observed. While such a procedure can be employed quite generally with small designs or with highly significant results, it is not feasible for hand computation in other cases. Fortunately the method readily adapts to automatic machine computation, and it would be easy to write programs for the routine calculation of exact significance probabilities.

[4] For Trimming and Winsorizing see Sect. 14 of "The future of data analysis" by John W. Tukey (1962). *Ann. Math. Statist.* **33** 1–67.

In the special case of the comparison of two treatments, enumeration by hand can be organized so that it is applicable in designs considerably larger than that considered in Section 3. Suppose that the subjects are divided into several blocks, that there are two treatments A and B applied with unrestricted randomization within blocks, and that the test statistic W is the sum of the ranks after alignment associated with, say, treatment A. (This is the situation illustrated by the example in Section 3.) Let n denote the number of blocks, let there be N_i subjects in the ith block of which s_i are allocated to the first treatment, and let W_i be the sum of the ranks after alignment of these s_i subjects. Then $W = W_1 + \cdots + W_n$.

Under the null hypothesis the W_i are independent random variables. The possible values of W_i are integers, and its null distribution can easily be written down by considering the $\binom{N_i}{s_i}$ equally likely choices of ranks of the first treatment. The calculation of the null distribution of W now requires the convolution of the k independent integer-valued random variables. There are $\prod \binom{N_i}{s_i}$ equally likely cases, and we have only to count those giving $W = w$ to find $P(W = w)$. A convenient layout for this count essentially like that required in the exact use of the Wilcoxon sum statistic W_s, may be presented through an example.

Suppose there are $k = 5$ blocks whose sizes, allocations, and ranks after alignment are as given in Table 4.1. The ranks of subjects receiving treatment A are underlined, and the observed value of W is 79, which is 22 greater than its smallest possible value 57. To compute $P(W \leqq 79)$ it will only be necessary to consider the lower tail for a range of 22 units at each step.

We first write down by inspection the distribution of $W_i - \min W_i = U_i$ for each block, multiplying each probability by $\binom{N_i}{s_i}$ so that only integers need be written. The results are shown in the top section of Table 4.2. For example, when two ranks are chosen from those in Block 1, the six equally likely values

TABLE 4.1

i	N_i	s_i	$\binom{N_i}{s_i}$	Ranks after alignment					W_i	min W_i
1	4	2	6	4	5	17	20		21	9
2	5	3	10	6	7	12	13	15	25	25
3	3	1	3	9	10	16			10	9
4	4	2	6	2	8	11	21		19	10
5	5	2	10	1	3	14	18	19	4	4
			$\prod \binom{N_i}{s_i} = 10800$						$W = 79$	min $W = 57$

TABLE 4.2

i	u = 0					5						10						15						20				Check Column
1	1														1	1			1									
2	1	1		1			1	1	1	2	1							1										
3	1	1						1																				
4	1			1						1						1							1				1	
5	1											1				1		1	1	1	1							
	1														1	1		1		1								5
		1														1	1		1									5
				1														1	1									5
							1												1	1	1			1		1		5
								1												1	1			1		1		4
									1														1	1				3
										2														2		2		6
											1															1		2
																		1										1
	1	1		1			1	1	1	2	1			1	2	1	3	1	3	1	2	3	2	4	5			36
		1	1		1			1	1	1	2	1			1	2	1	3	3	1	2	1	3	2	4			31
								1	1		1					1	1	2	1			1	2	1	3			16
	1	2	1	1	1		1	3	3	3	4	1	1	1	4	4	5	8	5	3	6	7	7	12			83
				1	2	1	1	1		1	3	3	1	3	4	1	1	4	4	5	8	5	3	6			57
										1	2	1	1	1	1		1	3	3	3	4	1	1	4			36
															1	2	1	1	1		1	3	3	3			16
																					1	2	1	1			5
																								1			1
	1	2	1	2	3	1	2	4	3	5	9	5	5	10	7	8	16	13	11	20	18	15	27				188
												1	2	1	2	3	1	2	4	3	5	9	5				38
														1	2	1	2	3	1	2	4	3	5				24
																1	2	1	2	3	1	2	4				16
																	1	2	1	2	3	1	2				12
																		1	2	1	2	3	1				10
																			1	2	1	2	3				9
	1	2	1	2	3	1	2	4	3	5	9	6	7	12	11	13	22	22	22	33	34	35	47				297
	1	3	4	6	9	10	12	16	19	24	33	39	46	58	69	82	104	126	148	181	215	250	297				

for W_1 are $4 + 5 = 9, 4 + 17 = 21, 5 + 17 = 22, 4 + 20 = 24, 5 + 20 = 25$ and $17 + 20 = 37$. Therefore the values of U_1 are $9 - 9 = 0$, $21 - 9 = 12$, $22 - 9 = 13$, $24 - 9 = 15$, $25 - 9 = 16$ and $37 - 9 = 28$. The last of these is outside the range of interest; the other five are represented in the row of Table 4.2 corresponding to $i = 1$.

The convolution of the five random variables U_i is displayed in the remainder of Table 4.2. The second section shows the convolution of U_1 and U_2. The rows correspond to values of U_2, the columns to values of $U_1 + U_2$. Thus, the first row of the second section gives for each possible value of $u_1 + u_2$ the number of cases for which $u_1 = 0$; the second row those for which $u_1 = 1$ (these are obtained from the first row by shifting it to the right by one unit); etc. By adding the number in each column, we obtain the distribution of $U_1 + U_2$ $\left(\text{with each probability multiplied by } \binom{N_1}{s_1}\binom{N_2}{s_2}\right)$, which is given in the top line of the third section. Similarly, the third section represents the convolution of $U_1 + U_2$ with U_3, etc. The final results are in the next to last line, so that $P(\sum U_i = 22) = P(W = 79) = 47/10800$. Cumulating, the last line gives $P(W \leqq 79) = 297/10800 = .0275$.

The computation requires only additions and proceeds quite rapidly. If desired a check is provided by the row sums as shown on the right. The method will be effective so long as the observed value w of W is not too much larger than its minimum value (or too much smaller than its maximum, if $P(W \geqq w)$ is to be computed).

In principle, the exact joint distribution of the rank sums could be similarly handled if there are more than two treatments, but the labor is much greater. For example, if there are treatments A, B, C, with rank sums W, X, Y, we may regard (W, X) as the sum of n independent vectors (W_i, X_i) each of which is integer-valued and has a distribution whose terms when multiplied by the multinomial coefficient $\binom{N_i}{s_i, t_i}$ are integers. However, each step in the convolution requires a double summation, so that the layout is less simple than that shown above.

5. Normal approximation to the null distribution of W. With designs much larger than that used for illustration above, exact hand computation becomes excessively lengthy. Fortunately, in such cases a simple normal approximation may serve. Recall that $W = W_1 + \cdots + W_n$ is the sum of n independent random variables W_i. Each W_i, as the sum of a sample drawn from a fixed set of integers, has easily calculated moments, so the low order moments of W may be readily obtained. The W_i will usually have variances of about the same magnitude; thus, if n is reasonably large one may expect W to be approximately normal. This hope is reinforced by a theorem proved in the next section.

Let us illustrate the method on the example of the preceding section. If the N_i ranks in Block i are denoted by $r_{ij}, j = 1, \cdots, N_i$, then $EW_i = s_i \sum r_{ij}/N_i$ while

TABLE 5.1

i	N_i	s_i	Ranks					EW_i	var W_i
1	4	2	4	5	17	20		23	67
2	5	3	6	7	12	13	15	31.8	18.36
3	3	1	9	10	16			11.67	9.56
4	4	2	2	8	11	21		21	63
5	5	2	1	3	14	18	19	22	85.8
								$E(W) = 109.47$	var$(W) = 243.72$ $\sigma(W) = 15.61$

TABLE 5.2

w	$P(W \leqq w)$	Normal approximation	Error
79	.0275	.0274	.0001
78	.0231	.0236	.0005
77	.0199	.0203	.0004
76	.0168	.0173	.0005
75	.0137	.0148	.0011
74	.0117	.0125	.0008

$$\text{var}\,(W_i) = \frac{s_i\, t_i}{N_i - 1}\left[(\sum r_{ij}^2)/N_i - (\sum r_{ij}/N_i)^2\right],$$

where in each case the summation is for j from 1 to N_i. Finally $EW = \sum EW_i$ and $\text{var}(W) = \sum \text{var}(W_i)$, as shown in Table 5.1. As W is integer-valued, it is natural to use a continuity correction. We approximate $P(W \leqq 79)$ by

$$\Phi\left(\frac{79.5 - 109.47}{15.61}\right) = \Phi(-1.9199) = .0274.$$

The excellent agreement of this value with the correct .0275 is something of an accident, as the following table shows. However, even with only 5 blocks the approximation would be accurate enough for most purposes (see Table 5.2).

We remark that the third and fourth moments of W are also easily available if it is thought desirable to use an approximation based on more than two moments.

6. Asymptotic null distribution of the blocked Wilcoxon statistic. In the preceding section we have discussed the exact null distribution of the blocked Wilcoxon statistic, as well as the normal approximation. In the present section we shall show that this distribution when normalized in the usual way tends to the standard normal distribution. In the present section, we shall consider the limiting behavior of this distribution as the number of blocks becomes large. As before, we shall consider the responses in each block (without regard as to which belongs

to treatment and which to control) as fixed, and we shall refer to the totality of these responses as the *configuration*. The only randomness is that resulting from the independent random assignments of treatments in each block. We shall prove below under certain assumptions that the null distribution of the blocked Wilcoxon statistic when normalized in the usual way tends to the standard normal distribution and that the convergence is *uniform* in the configurations.

This is proved by means of the Berry-Esseen theorem (cf., Theorem B on p. 288 of [8]) which may be stated as follows:

Let $W_1, W_2, \cdots$ be independently distributed, with means $\mu_1, \mu_2, \cdots$, and let

$$E(W_i - \mu_i)^2 = \sigma_i^2, \qquad E|W_i - \mu_i|^3 = \beta_i, \quad S_n^2 = \sum_{i=1}^{n} \sigma_i^2.$$

Let F_n denote the c.d.f. of $\sum_{i=1}^{n} (W_i - \mu_i)/S_n$, and Φ the standard normal c.d.f. There exists a constant $c < \infty$ such that for all x:

$$|F_n(x) - \Phi(x)| \leqq \frac{c}{S_n^3} \sum_{i=1}^{n} \beta_i . \tag{6.1}$$

We are concerned with the comparison of a treatment with a control in n blocks. As before, suppose that the ith block contains N_i experimental subjects, of which t_i are selected at random to receive the treatment, with the remaining s_i serving as controls ($s_i + t_i = N_i$, $i = 1, \cdots, n$). We shall assume that

$$N_i \leqq k \quad \text{for all} \quad i. \tag{6.2}$$

LEMMA 1. *Let W_i denote the sum of the ranks (after alignment) of the control responses in the ith block and let β_i denote the absolute* 3rd *moment of W_i. Then under assumption* (6.2) *there exists a constant $0 < b < \infty$ independent of the configuration such that*

$$\sum_{i=1}^{n} \beta_i \leqq bn^4. \tag{6.3}$$

PROOF. Note that

$$E|W_i - \mu_i|^3 \leqq s_i \max |r_{ij} - \mu_i|^3 \leqq k \cdot N^3,$$

where $r_{i1}, \cdots, r_{iN_i}$ denote the ranks in the ith block and where $N = \sum N_i$.

Since $N \leqq kn$, it follows that

$$\sum_{i=1}^{n} \beta_i \leqq n \cdot k^4 n^3 = k^4 n^4$$

as was to be proved.

We note that (6.3) is valid without any assumptions regarding the method of alignment, and that it does not require complete randomization within each block but would be equally valid under any method of restricted randomization. On the other hand, to obtain a lower bound for S_n^2 we make the assumption:

(A) After alignment, each block contains at least one observation above and one below the origin.

This assumption is satisfied for all reasonable methods of alignment such as alignment on the mean, on a censored or Winsorized mean, or on the median.

In addition, we shall suppose that:

(B) Within each block, complete randomization is employed.

LEMMA 2. *Let S_n^2 be the sum of the variances of the variables W_i of Lemma* 1. *Under assumption* (6.2), (A) *and* (B), *there exists a constant $0 < a < \infty$ independent of the configuration such that*

$$S_n^2 \geqq an^3. \tag{6.4}$$

PROOF. Under assumption (B), it is well known that

$$\sigma_i^2 = \frac{N_i - 1}{N_i - s_i} s_i \tau_i^2 \qquad \text{where} \qquad \tau_i^2 = \sum_{j=1}^{N_i} (r_{ij} - \bar{r}_i)^2/N_i .$$

Since this is a minimum when $s_i = 1$, in which case $\sigma_i^2 = \tau_i^2$, it is enough to prove that $\sum_{i=1}^n \tau_i^2 \geqq an^3$. Now

$$\tau_i^2 \geqq (1/N_i) \min_j (r_{ij} - \bar{r})^2 \geqq (1/k) \min_j (r_{ij} - \bar{r})^2$$

since of the r_{ij} at least one lies on either side of $\bar{r}$ and hence at least one must be closer to $\bar{r}$ than to $\bar{r}_i$. Since the N ranks r_{ij} are distinct integers, it follows that

$$\sum_i \min_j (r_{ij} - \bar{r})^2 \geqq 2\left(1^2 + 2^2 + \cdots + \left[\frac{n-2}{2}\right]^2\right)$$

and since the right-hand side is of the order n^3, this completes the proof.

THEOREM 2. *Under assumptions* (6.2), (A) *and* (B), *the distribution F_n of $\sum_{i=1}^n (W_i - \mu_i)/S_n$ tends to the standard normal distribution as $n \to \infty$, and the convergence is uniform in the configuration.*

PROOF. Combining (6.3) and (6.4) with (6.1), we see that

$$|F_n(x) - \Phi(x)| \leqq (bc/a^{\frac{1}{2}})(1/n^{\frac{1}{2}}),$$

which proves the desired result.

In order to apply (6.1) it is of course not necessary that (6.3) and (6.4) be satisfied but only that

$$\sum_{i=1}^n \beta_i/S_n^3 \to 0 \quad \text{as} \quad n \to \infty. \tag{6.5}$$

If for example, the ranks in the first block are $1, \cdots, N_1$; those in the second block $N_1 + 1, \cdots, N_1 + N_2$; etc., both S_n^2 and $\sum_{i=1}^n \beta_i$ are of order n, and hence (6.5) holds without (6.4) being satisfied.

On the other hand, although assumptions (A) and (B) are certainly not necessary for the validity of Theorem 2, some conditions on the method of alignment and randomization are required as the following two examples will show.

EXAMPLE 6.1. Suppose $N = 2n$ and the ith block consists of the ranks $(2i, 2i + 1)$ for $i = 1, \cdots, n - 1$ while the nth block consists of the ranks 1 and $2n$. Within each block, let one of the ranks be assigned at random to treatment, the other to control.

EXAMPLE 6.2. Suppose $N = 4n$ and the ranks in block 1 are $1, 2, 4n - 1, 4n$; those in block 2 are $3, 4, 4n - 3, 4n - 2$; etc., and suppose that in each block except the first one, the probability is $\frac{1}{2}$ that the two outside ranks belong to treatment and $\frac{1}{2}$ that the inner two ranks belong to treatment.

In the first of these examples, condition (B) is satisfied but not (A); in the second example, condition (A) is satisfied but not (B). For both examples, it is clear that $\sum W_i$ does not tend to be normally distributed as $n \to \infty$.

7. Efficiency. Some indications regarding the efficiency of the W-test discussed in the preceding sections can be obtained by considering the special case of paired comparisons $(s_i = t_i = 1)$.[5] Let W be the rank sum oe the treated subjects, after each block (of two) has been aligned on the midpoint between the two observations X (control) and Y (treated) in this block. To determine the efficiency of the W-test, we shall obtain a relation between W and the Wilcoxon one-sample statistic W'.

Suppose without loss of generality that the n blocks are numbered in order of increasing absolute difference $Y - X$ between treated and control response. Let

$$r_i = i, \qquad s_i = 0 \quad \text{if in the } i\text{th block} \quad X_i < Y_i,$$

$$r_i = 0, \qquad s_i = i \quad \text{otherwise.}$$

Further, let r_i^* denote the rank of the Y-observation in the ith block after alignment. Then if we have $X_i < Y_i$, clearly

$$r_i^* = n + i = n + r_i,$$

since after alignment Y_i exceeds all n negative observations and is the ith smallest of the positive observations.

Similarly, if $Y_i < X_i$, it is seen that

$$r_i^* = n + 1 - i = n + 1 - s_i.$$

The Wilcoxon statistic after alignment is then

$$W = \sum_{i=1}^{n} r_i^* = \sum^{(1)} (n + r_i) + \sum^{(2)} (n + 1 - s_i)$$

where $\sum^{(1)}$ and $\sum^{(2)}$ extend over all blocks with $Y > X$ and $Y < X$ respectively. If S is the number of blocks in which $Y > X$, that is, the sign statistic,

[5] In this case the efficiency computation is very much simpler than it is in general since the conditional null distribution of the test statistic given the configuration is independent of the configuration.

we have

$$W = \sum_{i=1}^{n} r_i + nS + (n+1)(n-S) - \sum_{i=1}^{n} s_i$$

$$= \sum_{i=1}^{n} r_i - \sum_{i=1}^{n} s_i - S + n(n+1).$$

Finally, since $\sum_{i=1}^{n} r_i + \sum_{i=1}^{n} s_i = 1 + 2 \cdots + n = \frac{1}{2}n(n+1)$, we get the relationship

$$W = 2W' - S + \tfrac{1}{2}[n(n+1)]$$

where $W' = \sum_{i=1}^{n} r_i$ is the one-sample Wilcoxon statistic.

Since the variance of W' is of order n^3 and the variance of S only of order n, it is seen that W and W' are asymptotically equivalent, and that the two associated tests have the same Pitman efficiency. In particular, the asymptotic efficiency of W relative to the corresponding t-test is $3/\pi$.

Suppose now that the block sizes are even but larger than 2 and that $s_i = t_i$ for all i. We can then obtain a test of asymptotic efficiency $3/\pi$ relative to the t-test by pairing control and treatment observations within each block at random and applying the W'-test to the resulting pairs. It seems plausible that an efficient method of alignment of the block as a whole followed by an application of the W'-test should be more efficient than this rather arbitrary procedure,[6] and preliminary work for the case of normal distributions with alignment on the block mean suggests that this is indeed the case.

This slight gain in efficiency appears however to decrease and tend to zero if instead of a large number of small blocks we are dealing with a small number of large blocks. As the block size tends to infinity, the additional information gained from intrablock comparisons above that provided by interblock comparisons, seems to tend to zero, with the efficiency of the W'-method tending to the efficiency $3/\pi$ found in Section 2 as the limiting efficiency for the Wilcoxon-van Elteren method based on independent rankings.

We hope to amplify these remarks, which are based partly on heuristic reasoning and partly on preliminary computations, in a later paper.

8. Several treatments. We conclude with two examples which indicate how the technique of aligned ranks may be applied when more than two treatments are to be compared.

EXAMPLE 8.1. *Incomplete block design.* Suppose three treatments are compared on three matched pairs, with these results

[6] Another interesting possibility would be to replace random pairing by a pairing of the smallest control with the smallest treatment observation, of the second smallest control with the second smallest treatment observation, etc., and then to compute the one-sample Wilcoxon statistic for these pairs.

	Treatment		
Block	A	B	C
1	131	115	—
2	—	151	141
3	131	—	105

If we align on block mean, the ranks are as follows

	Treatment		
Block	A	B	C
1	5	2	—
2	—	4	3
3	6	—	1

Let us use the Kruskal-Wallis statistic, which tests against unspecified differences in the treatments. The treatments A, B, C have rank sums 11, 6, 4 respectively, and the sum of squared differences from the mean of 7 is

$$(11 - 7)^2 + (6 - 7)^2 + (4 - 7)^2 = 26.$$

There are $2^3 = 8$ equally likely ways to label the ranks in the three blocks, and the actual labeling is seen by inspection to give the largest value of the statistic. Thus the significance probability is $\frac{1}{8}$.

EXAMPLE 8.2. *Latin squares.* To illustrate the application of our method to Latin squares, suppose that three treatments A, B, C are compared in two 3×3 squares, yielding these observations

B 4.461	C 2.798	A 7.402	C 5.424	B 9.670	A 9.669
A 3.412	B 2.405	C 5.227	B 5.062	A 9.368	C 5.710
C 3.454	A 2.169	B 6.717	A 6.605	C 7.786	B 7.427
	Square 1			Square 2	

When the row and column effects have been removed from each square in the usual way, these residuals are obtained:

B .025 (10)	C −.319 (5)	A .293 (13)
A .182 (12)	B .494 (15)	C −.676 (4)
C −.208 (6)	A −.174 (7)	B .382 (14)
	Square 1	
C −1.114 (2)	B −.112 (8)	A 1.226 (18)
B .065 (11)	A 1.127 (17)	C −1.192 (1)
A 1.049 (16)	C −1.014 (3)	B −.034 (9)
	Square 2	

After each residual is written its rank in the residual pool. The rank sums of treatments A, B, C are respectively

$$U = 32 + 51 = 83 \qquad V = 39 + 28 = 67 \qquad W = 15 + 6 = 21$$

According to the null hypothesis of no treatment effect, the 18 ranks are fixed numbers to which the labels A, B, C were attached in a pattern of restricted randomization. In a 3×3 square there are only two basic patterns, but in addi-

tion the labels may be arranged in $3! = 6$ orders, giving $2 \cdot 6 = 12$ cases in all. As each square has 12 cases, there are $12^2 = 144$ cases, which are equally likely in the usual design. Whatever test statistic may be used, its null distribution may be obtained by calculating its value for each of these cases.

For example, suppose we again use the Kruskal-Wallis statistic,

$$T = (U - 57)^2 + (V - 57)^2 + (W - 57)^2.$$

With the actual data, $T = 2072$. To determine the significance of this value, notice that in Square 1 the three rank sums must be either 15, 32, 39 or else 21, 31, 34, depending on which basic pattern is used. Similarly, in Square 2 we should have either 25, 28, 32 or 6, 28, 51. Now T is symmetric in U, V, W, which means that we may attach the labels A, B, C arbitrarily in one square. There are 24 equally likely values of T, according to the four choices of pattern and the six orderings in Square 2. Inspection shows that only one of these arrangements,

$$U = 32 + 28 = 60 \qquad V = 39 + 51 = 90 \qquad W = 15 + 6 = 21$$

yields a larger value of T than that actually observed. Thus the experiment has significance probability $2/24 = 0.083$.

If there are many blocks or squares, direct enumeration would have to be done on an automatic computing machine. Fortunately there is a relatively simple normal approximation. In Block i or Square i, the sums of ranks of treatments A and B gives a vector U_i, V_i whose early moments are easy to compute. From these the moments of $U = \sum U_i$, $V = \sum U_i$ may be obtained. If one uses as test statistic an appropriately chosen quadratic function of U and V, the null distribution would be approximately χ^2_2. This method readily extends to blocks or squares with more than three treatments.

REFERENCES

[1] BENARD, A. and ELTEREN, PH. VAN (1953). A generalization of the method of m-rankings. *Indag. Math.* **15** 358–369.

[2] DURBIN, J. (1951). Incomplete blocks in ranking experiments. *British J. Psych.* **4** 85–90.

[3] ELTEREN, PH. VAN (1960). On the combination of independent two sample tests of Wilcoxon. *Bull. Inst. Internat. Statist.* **37** 351–361.

[4] ELTEREN, PH. VAN and NOETHER, G. E. (1959). The asymptotic efficiency of the χ^2_n-test for a balanced incomplete block design. *Biometrika* **46** 475–477.

[5] FRIEDMAN, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32** 675–698.

[6] HODGES, J. L., JR. and LEHMANN, E. L. (1956). The efficiency of some non-parametric competitors of the t-test. *Ann. Math. Statist.* **27** 324–335.

[7] HODGES, J. L., JR. and LEHMANN, E. L. (1961). Comparison of the normal scores and Wilcoxon tests. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 307–317. Univ. of California Press.

[8] LOÈVE, M. (1960). *Probability Theory*, 2nd ed. Van Nostrand, Princeton.

[9] WALSH, JOHN E. (1959). Exact nonparametric tests for randomized blocks. *Ann. Math. Statist.* **30** 1034–1040.

[10] WILCOXON, F. (1946). Individual comparisons of grouped data by ranking methods. *J. of Entomology* **39** 269–270.