Use of Ranks in One-Criterion Variance Analysis

William H. Kruskal; W. Allen Wallis

Journal of the American Statistical Association, Vol. 47, No. 260. (Dec., 1952), pp. 583-621.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 260 DECEMBER 1952 *Volume 47*

USE OF RANKS IN ONE-CRITERION VARIANCE ANALYSIS

WILLIAM H. KRUSKAL AND W. ALLEN WALLIS
University of Chicago

Given C samples, with n_i observations in the ith sample, a test of the hypothesis that the samples are from the same population may be made by ranking the observations from from 1 to $\sum n_i$ (giving each observation in a group of ties the mean of the ranks tied for), finding the C sums of ranks, and computing a statistic H. Under the stated hypothesis, H is distributed approximately as $\chi^2(C-1)$, unless the samples are too small, in which case special approximations or exact tables are provided. One of the most important applications of the test is in detecting differences among the population means.*

1. INTRODUCTION

1.1. *Problem*

A COMMON problem in practical statistics is to decide whether several samples should be regarded as coming from the same population. Almost invariably the samples differ, and the question is whether the differences signify differences among the populations, or are merely the chance variations to be expected among random samples from the same population. When this problem arises one may often assume that the populations are of approximately the same form, in the sense that if they differ it is by a shift or translation.

1.2. *Usual Solution*

The usual technique for attacking such problems is the analysis of variance with a single criterion of classification [46, Chap. 10]. The variation among the sample means, $\bar{x}_i$, is used to estimate the variation among individuals, on the basis of (*i*) the assumption that the variation among the means reflects only random sampling from a population in which individuals vary, and (*ii*) the fact that the variance of the means of random samples of size n_i is σ^2/n_i where σ^2 is the population variance. This estimate of σ^2 based on the variation among sample means is then compared with another estimate based only on the varia-

* Based in part on research supported by the Office of Naval Research at the Statistical Research Center, University of Chicago.

For criticisms of a preliminary draft which have led to a number of improvements we are indebted to Maurice H. Belz (University of Melbourne), William G. Cochran (Johns Hopkins University), J. Durbin (London School of Economics), Churchill Eisenhart (Bureau of Standards), Wassily Hoeffding (University of North Carolina), Harold Hotelling (University of North Carolina), Howard L. Jones (Illinois Bell Telephone Company), Erich L. Lehmann (University of California), William G. Madow (University of Illinois), Henry B. Mann (Ohio State University), Alexander M. Mood (The Rand Corporation), Lincoln E. Moses (Stanford University), Frederick Mosteller (Harvard University), David L. Russell (Bowdoin College), I. Richard Savage (Bureau of Standards), Frederick F. Stephan (Princeton University), Alan Stuart (London School of Economics), T. J. Terpstra (Mathematical Center, Amsterdam), John W. Tukey (Princeton University), Frank Wilcoxon (American Cyanamid Company), and C. Ashley Wright (Standard Oil Company of New Jersey), and to our colleagues K. A. Brownlee, Herbert T. David, Milton Friedman, Leo A. Goodman, Ulf Grenander, Joseph L. Hodges, Harry V. Roberts, Murray Rosenblatt, Leonard J. Savage, and Charles M. Stein.

tion within samples. The agreement between these two estimates is tested by the variance ratio distribution with $C-1$ and $N-C$ degrees of freedom (where N is the number of observations in all C samples combined), using the test statistic $F(C-1, N-C)$. A value of F larger than would ordinarily result from two independent sample estimates of a single population variance is regarded as contradicting the hypothesis that the variation among the sample means is due solely to random sampling from a population whose individuals vary.

When σ^2 is known, it is used in place of the estimate based on the variation within samples, and the test is based on the $\chi^2(C-1)$ distribution (that is, χ^2 with $C-1$ degrees of freedom) using the test statistic

$$\chi^2(C-1) = \sum \frac{(\bar{x}_i - \bar{x})^2}{\sigma^2/n_i} \tag{1.1}$$

where $\bar{x}$ is the mean of all N observations.

1.3. *Advantages of Ranks*

Sometimes it is advantageous in statistical analysis to use ranks instead of the original observations—that is, to array the N observations in order of magnitude and replace the smallest by 1, the next-to-smallest by 2, and so on, the largest being replaced by N. The advantages are:

(1) The calculations are simplified. Most of the labor when using ranks is in making the ranking itself, and short cuts can be devised for this. For example, class intervals can be set up as for a frequency distribution, and actual observations entered instead of tally marks. Another method is to record the observations on cards or plastic chips[1] which can be arranged in order, the cards perhaps by sorting devices.

(2) Only very general assumptions are made about the kind of distributions from which the observations come. The only assumptions underlying the use of ranks made in this paper are that the observations are all independent, that all those within a given sample come from a single population, and that the C populations are of approximately the same form. The F and χ^2 tests described in the preceding section assume approximate normality in addition.

(3) Data available only in ordinal form may often be used.

(4) When the assumptions of the usual test procedure are too far from reality, not only is there a problem of distribution theory if the usual test is used, but it is possible that the usual test may not have as good a chance as a rank test of detecting the kinds of difference of real interest.

The present paper presents an analog, based on ranks and called the H test, to one-criterion variance analysis.

[1]We are indebted to Frank Wilcoxon for this suggestion.

1.4. *The H Test*

The rank test presented here requires that all the observations be ranked together, and the sum of the ranks obtained for each sample. The test statistic to be computed if there are no ties (that is, if no two observations are equal) is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{C} \frac{R_i^2}{n_i} - 3(N+1) \qquad \textit{(no ties)} \tag{1.2}$$

where

C = the number of samples,
n_i = the number of observations in the ith sample,
$N = \sum n_i$, the number of observations in all samples combined,
R_i = the sum of the ranks in the ith sample.

Large values of H lead to rejection of the null hypothesis.

If the samples come from identical continuous populations and the n_i are not too small, H is distributed as $\chi^2(C-1)$, permitting use of the readily available tables of χ^2. When the n_i are small and $C=2$, tables are available which are described in Section 5.3. For $C=3$ and all $n_i \leqq 5$, tables are presented in Section 6. For other cases where the χ^2 approximation is not adequate, two special approximations are described in Section 6.2.

If there are ties, each observation is given the mean of the ranks for which it is tied. H as computed from (1.2) is then divided by

$$1 - \frac{\sum T}{N^3 - N} \tag{1.3}$$

where the summation is over all groups of ties and $T = (t-1)t(t+1) = t^3 - t$ for each group of ties, t being the number of tied observations in the group. Values of T for t up to 10 are shown in Table 1.1.[2]

TABLE 1.1

(See Section 3.1.2)

t	1	2	3	4	5	6	7	8	9	10
T	0	6	24	60	120	210	336	504	720	990

Since (1.3) must lie between zero and one, it increases (1.2). If all N observations are equal, (1.3) reduces (1.2) to the indeterminate form 0/0. If there are no ties, each value of t is 1 so $\sum T = 0$ and (1.2) is

[2] DuBois [4, Table I] gives values of T/12 (his c_t) and T/6 (his c_t) for t (his N) from 5 to 50.

unaltered by (1.3). Thus, (1.2) divided by (1.3) gives a general expression which holds whether or not there are ties, assuming that such ties as occur are given mean ranks:

$$H = \frac{\dfrac{12}{N(N+1)}\sum_{i=1}^{C}\frac{R_i^2}{n_i} - 3(N+1)}{1 - \sum T/(N^3 - N)}. \tag{1.4}$$

In many situations the difference between (1.4) and (1.2) is negligible. A working guide is that with ten or fewer samples a χ^2 probability of 0.01 or more obtained from (1.2) will not be changed by more than ten per cent by using (1.4), provided that not more than one-fourth of the observations are involved in ties.[3] H for large samples is still distributed as $\chi^2(C-1)$ when ties are handled by mean ranks; but the tables for small samples, while still useful, are no longer exact.

For understanding the nature of H, a better formulation of (1.2) is

$$H = \frac{N-1}{N}\sum_{i=1}^{C}\frac{n_i[\overline{R}_i - \tfrac{1}{2}(N+1)]^2}{(N^2-1)/12} \qquad (no\ ties) \tag{1.5}$$

where $\overline{R}_i$ is the mean of the n_i ranks in the ith sample. If we ignore the factor $(N-1)/N$, and note that $\frac{1}{2}(N+1)$ is the mean and $\frac{1}{12}(N^2-1)$ the variance of the uniform distribution over the first N integers, we see that (1.5), like (1.1), is essentially a sum of squared standardized deviations of random variables from their population mean. In this respect, H is similar to χ^2, which is defined as a sum of squares of standardized *normal* deviates, subject to certain conditions on the relations among the terms of the sum. If the n_i are not too small, the $\overline{R}_i$ jointly will be approximately normally distributed and the relations among them will meet the χ^2 conditions.

2. EXAMPLES

2.1 *Without Ties*

In a factory, three machines turn out large numbers of bottle caps. One machine is standard and two have been modified in different ways, but otherwise the machines and their operating conditions are identical. On any one day, only one machine is operated. Table 2.1

[3] Actually, for the case described it is possible for the discrepancy slightly to exceed ten per cent. For a given total number of ties, S, the second term of (1.3) is a maximum if all S ties are in one group and this maximum, $(S^3-S)/(N^3-N)$, is slightly less than $(S/N)^3$. Thus, for $S/N = \frac{1}{4}$, (1.3) $>63/64$. The 0.01 level of $\chi^2(9)$ is 21.666. This divided by 63/64 is 22.010, for which the probability is 0.00885, a change of $11\frac{1}{2}$ per cent. For higher probability levels, fewer samples, or more than one group of ties, the percentage change in probability would be less. With the S ties divided into G groups, the second term of (1.3) is always less than $[(S-h)^3+4h]/N^3$, where $h = 2(G-1)$.

shows the production of the machines on various days and the calculation of H as 5.656. The true probability, if the machines really are the same with respect to output, that H should be as large as 5.656 is shown in Figure 6.1 and Table 6.1 as 0.049. The approximation to this probability given by the $\chi^2(2)$ distribution is 0.059. Two more-complicated approximations described in Section 6.2 give 0.044 and 0.045.

TABLE 2.1

DAILY BOTTLE-CAP PRODUCTION OF THREE MACHINES.
(Artificial data.)

Standard		Modification 1		Modification 2		
Output	Rank	Output	Rank	Output	Rank	
340	5	339	4	347	10	
345	9	333	2	343	7	
330	1	344	8	349	11	
342	6			355	12	
338	3					
						Sum
n	5		3		4	12
R	24		14		40	78
R^2/n	115.2		65.333		400.	580.533

Checks: $\sum n = N = 12$ $\qquad$ $\sum R = \frac{1}{2}N(N+1) = 78$

$$H = \frac{12 \times 580.533}{12 \times 13} - 3 \times 13 = 5.656 \simeq \chi^2(2) \qquad \textit{from (1.2)}$$

$\Pr[\chi^2(2) \geqq 5.656] = 0.059$ $\qquad$ *from [9] or [13]*

$\Pr[H(5, 4, 3) \geqq 5.656] = 0.049$ $\qquad$ *from Table 6.1*

If the production data of Table 2.1 are compared by the conventional analysis of variance, $F(2, 9)$ is 4.2284, corresponding to a probability of 0.051.

2.2 *With Ties*

Snedecor's data on the birth weight of pigs [46, Table 10.12] are shown in Table 2.2, together with the calculation of H adjusted for the mean-rank method of handling ties. Here H as adjusted[4] is 18.566. The true probability in this case would be difficult to find, but the

[4] Note that, as will often be true in practice, the adjustment is not worth the trouble even in this case: by changing H from 18.464 to 18.566, it changed the probability by only 0.0003, or 3 per cent. Since there are 47 ties in 13 groups, we see from the last sentence of note 3 that (1.3) cannot be less than $1-(23^3+96)/56^3$, which is 0.9302.

TABLE 2.2

BIRTH WEIGHTS (lbs.) OF EIGHT LITTERS OF PIGS

Source: Snedecor [46], Table 10.12

Litter 1		Litter 2		Litter 3		Litter 4		Litter 5		Litter 6		Litter 7		Litter 8		
Wt.	Rank	Wt.	Rank	Wt.	Rank	Wt.	Rank	Wt.	Rank	Wt.	Rank	Wt.	Rank	Wt.	Rank	
2.0	8½	3.5	52½	3.3	47½	3.2	41	2.6	23	3.1	36	2.6	23	2.5	18½	
2.8	27½	2.8	27½	3.6	54½	3.3	47½	2.6	23	2.9	31½	2.2	12½	2.4	15½	
3.3	47½	3.2	41	2.6	23	3.2	41	2.9	31½	3.1	36	2.2	12½	3.0	34	
3.2	41	3.5	52½	3.1	36	2.9	31½	2.0	8½	2.5	18½	2.5	18½	1.4	4	
4.4	56	2.3	14	3.2	41	3.3	47½	2.0	8½			1.2	2½			
3.6	54½	2.4	15½	3.3	47½	2.5	18½	2.1	11			1.2	2½			
1.9	6	2.0	8½	2.9	31½	2.6	23									
3.3	47½	1.6	5	3.4	51	2.8	27½									
2.8	27½			3.2	41											
1.1	1			3.2	41											
n	10		8		10		8		6		4		6		4	56
R	317		216½		414		277½		105½		122		71½		72	1596
R^2/n	10,048.9		5859.031		17,139.6		9625.781		1855.042		3721.0		852.042		1296.0	50,397.396

Checks $\sum n = N = 56$ $\sum R = \frac{1}{2}N(N+1) = 1596$

$$\frac{12 \times 50397.396}{56 \times 57} - 3 \times 57 = 18.464 = H \text{ prior to mean-rank adjustment}$$ *from (1.2)*

Adjustment for mean ranks: *from Table 1.1*

t:	2	4	2	2	4	5	4	4	3	7	6	2	2
T:	6	60	6	6	60	120	60	60	24	336	210	6	6

$\sum T = 960$ $\qquad 1 - \frac{\sum T}{N^3 - N} = 1 - \frac{960}{175560} = 0.9945$ *from (1.3)*

$H = \frac{18.464}{0.9945} = 18.566 \simeq \chi^2(7)$ *from (1.4)*

$\Pr[\chi^2(7) \geqq 18.566] = 0.010$ *from [9] or [13]*

$\chi^2(7)$ approximation gives a probability of 0.010. Two better approximations described in Section 6.2 give 0.006 and 0.005.

The conventional analysis of variance [46, Sec. 10.8] gives $F(7, 48) = 2.987$, corresponding with a probability of 0.011.

3. JUSTIFICATION OF THE METHOD

3.1. *Two Samples*

The rationale of the H test can be seen most easily by considering the case of only two samples, of sizes n and $N-n$. As is explained in Section 5.3, the H test for two samples is essentially the same as a test published by Wilcoxon [61] in 1945 and later by others.

In this case, we consider either one of the two samples, presumably the smaller for simplicity, and denote its size by n and its sum of ranks by R. We ask whether the mean rank of this sample is larger (or smaller) than would be expected if n of the integers 1 through N were selected at random without replacement.

The sum of the first N integers is $\frac{1}{2}N(N+1)$ and the sum of their squares is $\frac{1}{6}N(N+1)(2N+1)$. It follows that the mean and variance of the first N integers are $\frac{1}{2}(N+1)$ and $\frac{1}{12}(N^2-1)$.

The means of samples of n drawn at random without replacement from the N integers will be normally distributed to an approximation close enough for practical purposes, provided that n and $N-n$ are not too small. The mean of a distribution of sample means is, of course, the mean of the original distribution; and the variance of a distribution of sample means is $(\sigma^2/n)[(N-n)/(N-1)]$, where σ^2 is the population variance, N is the population size, and n is the sample size. In this case, $\sigma^2 = \frac{1}{12}(N^2-1)$, so

$$\sigma_{\bar{R}}^2 = \frac{(N^2-1)(N-n)}{12n(N-1)} = \frac{(N+1)(N-n)}{12n} \tag{3.1}$$

where $\sigma_{\bar{R}}^2$ represents the variance of the mean of n numbers drawn at random without replacement from N consecutive integers. Letting $\bar{R}$ denote the mean rank for a sample of n,

$$\frac{\bar{R} - \frac{1}{2}(N+1)}{\sqrt{(N+1)(N-n)/12n}} \tag{3.2}$$

may be regarded as approximately a unit normal deviate. The square of (3.2) is H as given by (1.2) with[5] $C=2$, and the square of a unit normal deviate has the $\chi^2(1)$ distribution.

[5] This may be verified by replacing $\bar{R}$ in (3.2) by R/n and letting the two values of R_i in (1.2) be R and $\frac{1}{2}N(N+1)-R$, with n and $N-n$ the corresponding values of n_i.

Notice that this expression is the same, except for sign, whichever of the two samples is used to compute it. For if the first sample contains n ranks whose mean is $\overline{R}$, the other sample must contain $N-n$ ranks whose mean is

$$\frac{\frac{1}{2}N(N+1) - n\overline{R}}{N-n} \tag{3.3}$$

and the value of (3.2) is changed only in sign if we interchange n and $N-n$, and replace $\overline{R}$ by (3.3).

In the two-sample case the normal deviate is perhaps a little simpler to compute than is H; furthermore, the sign of the normal deviate is needed if a one-tail test is required. For computations, formula (3.2) may be rewritten

$$\frac{2R - n(N+1)}{\sqrt{n(N+1)(N-n)/3}}. \tag{3.4}$$

The null hypothesis is that the two samples come from the same population. The alternative hypothesis is that the samples come from populations of approximately the same form, but shifted or translated with respect to each other. If we are concerned with the one-sided alternative that the population producing the sample to which R and n relate is shifted upward, then we reject when (3.4) is too large. The critical level of (3.4) at the α level of significance is approximately K_α, the unit normal deviate exceeded with probability α, as defined by

$$\frac{1}{\sqrt{2\pi}}\int_{K_\alpha}^{\infty} e^{-\frac{1}{2}x^2}dx = \alpha. \tag{3.5}$$

Values of (3.4) as large as K_α or larger result in rejection of the null hypothesis. If the alternative is one-sided but for a downward shift, the null hypothesis is rejected when (3.4) is as small as $-K_\alpha$ or smaller. If the alternative is two-sided and symmetrical, the null hypothesis is rejected if (3.4) falls outside the range $-K_{\frac{1}{2}\alpha}$ to $+K_{\frac{1}{2}\alpha}$.

3.1.1. *Continuity adjustment.* It seems reasonable to expect that a continuity adjustment may be desirable, to allow for the fact that R, the sum of the ranks in one sample, can take only integral values, whereas the normal distribution is continuous.[6] In testing against a two-sided alternative to the null hypothesis, the adjustment is made

[6] An extensive comparison of exact probabilities for the two-sample test [28] with those based on the normal approximation indicates that the normal approximation is usually better with the continuity adjustment when the probability is above 0.02, and better without it when the probability is 0.02 or below. This comparison was made for us by Jack Karush, who has also rendered invaluable assistance with numerous other matters in the preparation of this paper.

by increasing or decreasing R by $\frac{1}{2}$, whichever brings it closer to $\frac{1}{2}n(N+1)$, before substituting into (3.4). (If $R=\frac{1}{2}n(N+1)$, ignore the continuity adjustment.) With a one-sided alternative, R is increased (decreased) by $\frac{1}{2}$ if the alternative is that the sample for which R is computed comes from the population which is to the left (right) of the other.

3.1.2. *Ties.* If some of the N observations are equal, we suggest that each member of a group of ties be given the mean of the ranks tied for in that group. This does not affect the mean rank, $\frac{1}{2}(N+1)$. It does, however, reduce the variance below $\frac{1}{12}(N^2-1)$. Letting $T=(t-1)t(t+1)$ for each group of ties, where t is the number of tied observations in the group, and letting $\sum T$ represent the sum of the values of T for all groups of ties, we have, instead of (3.1),

$$\sigma_{\bar{R}}^2 = \frac{N(N^2-1)-\sum T}{12Nn} \cdot \frac{N-n}{N-1} \tag{3.6}$$

as the variance of the mean rank for samples of n. When there are no ties, $\sum T=0$ and (3.6) reduces to (3.1), so (3.6) may be regarded as the general expression for $\sigma_{\bar{R}}^2$ when the mean-rank method is used for such ties as occur. Notice that (3.6) is the product of (3.1) and (1.3).

This adjustment comes about as follows:[7] The variance $\frac{1}{12}(N^2-1)$ is obtained by subtracting the square of the mean from the mean of the squares of N consecutive integers. If each of the t integers $(x+1)$ through $(x+t)$ is replaced by $x+\frac{1}{2}(t+1)$, the sum is not changed but the sum of the squares is reduced by

$$\sum_{i=1}^{t}(x+i)^2 - t\left(x+\frac{t+1}{2}\right)^2 = \frac{(t-1)t(t+1)}{12} = \frac{T}{12}. \tag{3.7}$$

So the mean of the squares, and consequently the variance, is reduced by $T/12N$.

The mean-rank method of handling ties somewhat complicates the continuity adjustment, for the possible values of R are no longer simply the consecutive integers $\frac{1}{2}n(n+1)$ to $\frac{1}{2}n(2N-n+1)$, nor need they be symmetrical about $\frac{1}{2}(N+1)$. Our guess, however, is that it is better to make the $\pm\frac{1}{2}$ adjustment of Section 3.1.1. than not to make any.

---

[7] This is the adjustment alluded to by Friedman [10, footnote 11]. An equivalent adjustment for mean ranks has been suggested by Hemelrijk [16, formula (6)], but in a very complex form. A much simpler version of his formula (6) is obtained by multiplying our (3.6) by n^2. The same adjustment has been suggested by Horn [18a].

This adjustment, however, goes back at least as far as a 1921 paper by 'Student' [48a], applying it to the Spearman rank correlation coefficient. For further discussion and other references, see Kendall [20, Chap. 3].

An alternative method of handling ties is to assign the ranks at random within a group of tied observations. The distribution of H under the null hypothesis is then the same as if there had been no ties, since the null hypothesis is that the ranks are distributed at random. In order to use this method, adequate randomization must be provided with consequent complications in making and verifying computations. Some statisticians argue further that the introduction of extraneous random variability tends to reduce the power of a test. We do not know whether for the H test random ranking of ties gives more or less power than mean ranks; indeed, it may be that the answer varies from one alternative hypothesis to another and from one significance level to another.[8] When all members of a group of ties fall within the same sample, every assignment of their ranks gives rise to the same value of H, so that it might be thought artificial in this instance to use mean-ranks; even here, however, an eristic argument can be made for mean ranks, on the ground that H interprets a particular assignment of ranks against the background of all possible assignments of the same ranks to samples of the given sizes, and some of the possible assignments put the ties into different samples.[9]

3.1.3. *Examples.* (*i*) As a first example consider a particularly simple one discussed by Pitman [41].

TABLE 3.1

PITMAN EXAMPLE [41, p. 122]

Sample A		Sample B	
Observation	Rank	Observation	Rank
0	1	16	4
11	2	19	5
12	3	22	7
20	6	24	8
		29	9

$n=4, \quad N=9, \quad R=12$

[8] A few computations for simple distributions and small samples, some carried out by Howard L. Jones and some by us, showed mean ranks superior sometimes and random ranks others. For theoretical purposes, random ranking of ties is much easier to handle.For practical purposes, it should be remembered that there will ordinarily be little difference between the two methods; see notes 3 and 4. Computational considerations, therefore, lead us to suggest the mean-rank method.

Ranking of tied observations at random should be distinguished from increasing the power of a test by rejecting or accepting the null hypothesis on the basis of an ancillary random device, in such a way as to attain a nominal significance level which, because of discontinuities, could not otherwise be attained. Discussions of this are given by Eudey [6] and E. S. Pearson [37].

[9] This is illustrated in the calculation of the exact probability for the data of Table 3.2.

If we use the two-tail H test without adjustment for continuity, we compute the approximate unit-normal deviate from (3.4):

$$\frac{(2 \times 12) - (4 \times 10)}{\sqrt{(4 \times 10 \times 5)/3}} = -\frac{16}{\sqrt{200/3}} = -1.9596 \quad (\textit{no adjustment})$$

corresponding to a two-tail normal probability of 0.0500.

If we make the continuity adjustment, we get:

$$\frac{(2 \times 12.5) - (4 \times 10)}{\sqrt{(4 \times 10 \times 5)/3}} = -\frac{15}{\sqrt{200/3}} = -1.8371 \quad (\textit{continuity adjustment})$$

corresponding to a two-tail normal probability of 0.0662.

Actually, since the samples are so small, it is easy to compute the true probability under the null hypothesis of a value of R as extreme as, or more extreme than, 12. There are 9!/4!5! or 126 ways of selecting four ranks from among the nine, and all 126 ways are equally probable under the null hypothesis. Only four of the 126 lead to values of R of 12 or less. By symmetry another set of 4 lead to values as extreme but in the opposite direction, that is, $n(N+1)-R=28$ or more. Hence the true probability to compare with the foregoing approximations is 8/126, or 0.06349. This value can also be obtained from the tables given by Mann and Whitney [28]; they show 0.032 for one tail, and when doubled this agrees, except for rounding, with our calculation.[10]

(*ii*) A second, and more realistic, example will illustrate the kind of

TABLE 3.2

BROWNLEE EXAMPLE [2, p. 36]

Method A		Method B	
Value	Rank	Value	Rank
95.6	9½	93.3	4
94.9	7	92.1	3
96.2	12	94.7	5½
95.1	8	90.1	2
95.8	11	95.6	9½
96.3	13	90.0	1
		94.7	5½

$R = 60\frac{1}{2}$, $n = 6$, $N = 13$

[10] Pitman [41] gives a test which is like H except that it considers possible permutations of the actual observations instead of their ranks. For the example of Table 3.1, Pitman's test yields a two-tail probability of 5/126 or 0.03968.

complication that arises in practice. Table 3.2 shows the results of two alternative methods of technical chemical analysis. Since there are ties (two groups of two ties), mean ranks are used.

If we use (3.4) without adjusting either for continuity or for the use of mean ranks, we obtain as our approximate unit-normal deviate

$$\frac{121 - 84}{\sqrt{(84 \times 7)/3}} = \frac{37}{14} = 2.6429 \qquad (\textit{no adjustments})$$

which corresponds to the two-tail normal probability of 0.0082.

If we use the adjustment for mean ranks, we find that $\sum T = 12$, so (3.6) gives $\sigma_{\overline{R}} = 1.1635$ and the denominator of (3.4), which is

$$\sigma_{2R} = 2n\sigma_{\overline{R}}, \tag{3.8}$$

is adjusted to 13.9615. This leads to the approximate unit-normal deviate

$$\frac{121 - 84}{13.9615} = 2.6501 \qquad (\textit{adjusted for mean ranks})$$

corresponding to a two-tail probability of 0.0080—not appreciably different from the result without the adjustment.

The continuity adjustment is not desirable in this case, since the probability level is appreciably less than 0.02.[6] The comments of Section 3.1.2 about irregularities in the sequence of possible values of R also apply. For purely illustrative purposes, however, we note that the effect of the continuity adjustment would be to reduce R from $60\frac{1}{2}$ to 60, resulting in an approximate normal deviate of

$$\frac{120 - 84}{13.9615} = 2.5785 \qquad (\textit{adjusted for continuity and mean ranks})$$

for which the symmetrical two-tail normal probability is 0.0099.

The true probability in this case can be computed by considering all possible sets of six that could be selected from the 13 ranks 1, 2, 3, 4, $5\frac{1}{2}$, $5\frac{1}{2}$, 7, 8, $9\frac{1}{2}$, $9\frac{1}{2}$, 11, 12, 13. There are 13!/6!7! or 1716 such sets, all equally probable under the null hypothesis. Six of them give rise to values of R greater than or equal to $60\frac{1}{2}$, and five give rise to values of R less than or equal to $23\frac{1}{2}$, which is as far below as $60\frac{1}{2}$ is above $\frac{1}{2}n(N+1)$. Hence the true probability is 11/1716, or 0.00641.

3.2. *Three Samples*

When there are three samples, we may consider the average ranks for any two of them, say the ith and jth. The other sample, the kth,

would not tell us anything we cannot find out from two, for its mean rank must be

$$\overline{R}_k = \frac{\frac{1}{2}N(N+1) - (n_i\overline{R}_i + n_j\overline{R}_j)}{N - (n_i + n_j)}. \tag{3.9}$$

If the n's are not too small, the joint distribution of $\overline{R}_i$ and $\overline{R}_j$ will be approximately that bivariate normal distribution whose exponent is

$$-\frac{1}{2(1-\rho^2)}\left[\frac{\left(\overline{R}_i - \frac{N+1}{2}\right)^2}{\sigma_{\overline{R}_i}^2} - 2\rho\frac{\left(\overline{R}_i - \frac{N+1}{2}\right)\left(\overline{R}_j - \frac{N+1}{2}\right)}{\sigma_{\overline{R}_i}\sigma_{\overline{R}_j}} + \frac{\left(\overline{R}_j - \frac{N+1}{2}\right)^2}{\sigma_{\overline{R}_j}^2}\right]. \tag{3.10}$$

The variances needed in (3.10) are given by (3.1) and the correlation by

$$\rho = -\sqrt{\left(\frac{n_i}{N-n_i}\right)\left(\frac{n_j}{N-n_j}\right)} \tag{3.11}$$

which is the correlation between the means of samples of sizes n_i and n_j when all n_i+n_j are drawn at random without replacement from a population of N elements.[11] Thus the exponent (3.10) of the bivariate normal distribution which approximates the joint distribution of $\overline{R}_i$ and $\overline{R}_j$ is, when multiplied by -2,

$$\frac{12n_in_j}{N(N+1)(N-n_i-n_j)}\left[\left(\frac{N-n_j}{n_j}\right)\left(\overline{R}_i - \frac{N+1}{2}\right)^2 + 2\left(\overline{R}_i - \frac{N+1}{2}\right)\left(\overline{R}_j - \frac{N+1}{2}\right) + \left(\frac{N-n_i}{n_i}\right)\left(\overline{R}_j - \frac{N+1}{2}\right)^2\right]. \tag{3.12}$$

It is well known that -2 times the exponent of a bivariate normal dis-

[11] Although (3.11) is easily derived and is undoubtedly familiar to experts on sampling from finite populations, we have not found it in any of the standard treatises. It is a special case of a formula used by Neyman [47, p. 39] in 1923, and a more general case of one used by K. Pearson [38] in 1924. For assistance in trying to locate previous publications of (3.11) we are indebted to Churchill Eisenhart, Tore Dalenius (Stockholm), W. Edwards Deming (Bureau of the Budget), P. M. Grundy (Rothamsted Experimental Station) who told us of [38], Morris H. Hansen (Bureau of the Census), Maurice G. Kendall (London School of Economics), Jerzy Neyman (University of California) who told us of [47], June H. Roberts (Chicago), Frederick F. Stephan who provided a compact derivation of his own, John W. Tukey, and Frank Yates (Rothamsted Experimental Station).

tribution has the $\chi^2(2)$ distribution [32, Sec. 10.10]. Hence (3.12) could be taken as our test statistic for the three-sample problem, and approximate probabilities found from the χ^2 tables.

From the relations

$$n_i\overline{R}_i + n_j\overline{R}_j + n_k\overline{R}_k = \tfrac{1}{2}N(N+1) \tag{3.13}$$

and

$$n_i + n_j + n_k = N \tag{3.14}$$

it can be shown that the value of (3.12) will be the same whichever pair of samples is used in it, and that this value will be H as given by (1.2) with $C=3$. For computing, (1.2) has the advantages of being simpler than (3.12) and of treating all (R, n) pairs alike.

With three or more samples, adjustments for continuity are unimportant except when the n_i are so small that special tables of the true distribution should be used anyway.

Since the adjustment for the mean-rank method of handling ties is a correction to the sum of squares of the N ranks, it is the same for three or more groups as for two. The variances given by (3.1) for the case without ties are replaced by (3.6) when there are ties; hence (1.2) with mean ranks should be divided by (1.3) to give H as shown by (1.4).

3.3. *More than Three Samples*

Nothing essentially new is involved when there are more than three samples. If there are C samples, the mean ranks for any $C-1$ of them are jointly distributed approximately according to a multivariate normal distribution, provided that the sample sizes are not too small. The exponent of this $(C-1)$-variate normal distribution will have the same value whichever set of $C-1$ samples is used. This value, when multiplied by -2, will be H as given by (1.2), and it will be distributed approximately as $\chi^2(C-1)$, provided the n_i are not too small. The exponent of the approximating multivariate normal distribution is more complicated than for three samples, but it involves only the variances of the $\overline{R}_i$ as given by (3.6) and the correlations among pairs $(\overline{R}_i, \overline{R}_j)$ as given by (3.11).

By using matrix algebra, the general formula for H is obtained quite as readily as the formulas for two and three samples by the methods used in this paper. A mathematically rigorous discussion of H for the general case of C samples is presented elsewhere by Kruskal [25], together with a formal proof that its distribution under the null hypothesis is asymptotically χ^2.

4. INTERPRETATION OF THE TEST

4.1. *General Considerations*

H tests the null hypothesis that the samples all come from identical populations. In practice, it will frequently be interpreted, as is F in the analysis of variance, as a test that the population means are equal against the alternative that at least one differs. So to interpret it, however, is to imply something about the kinds of differences among the populations which, if present, will probably lead to a significant value of H, and the kinds which, even if present, will probably not lead to a significant value of H. To justify this or any similar interpretation, we need to know something about the power of the test: For what alternatives to identity of the populations will the test probably lead to rejection, and for what alternatives will it probably lead to acceptance of the null hypothesis that the populations are identical? Unfortunately, for the H test as for many nonparametric tests the power is difficult to investigate and little is yet known about it.

It must be recognized that relations among ranks need not conform to the corresponding relations among the data before ranking. It is possible, for example, that if an observation is drawn at random from each of two populations, the one from the first population is larger in most pairs, but the average of those from the second population is larger. In such a case the first population may be said to have the higher average rank but the lower average value.

It has been shown by Kruskal [25] that a necessary and sufficient condition for the H test to be consistent[12] is that there be at least one of the populations for which the limiting probability is not one-half that a random observation from this population is greater than an independent random member of the N sample observations. Thus, what H really tests is a tendency for observations in at least one of the populations to be larger (or smaller) than all the observations together, when paired randomly. In many cases, this is practically equivalent to the mean of at least one population differing from the others.

4.2. *Comparison of Means when Variability Differs*

Rigorously interpreted, all we can conclude from a significant value of H is that the populations differ, not necessarily that the means differ. In particular, if the populations differ in variability we cannot,

[12] A test is *consistent* against an alternative if, when applied at the same level of significance for increasing sample size, the probability of rejecting the null hypothesis when the alternative is true approaches unity. Actually, the necessary and sufficient condition stated here must be qualified in a way that is not likely to affect the interpretation of the H test suggested in this paragraph. An exact statement is given in [25].

strictly speaking, infer from a significant value of H that the means differ. In the data of Table 3.2, for example, the variances of the two chemical methods differ significantly (normal theory probability less than 0.01) and substantially (by a factor of 16), as Brownlee shows [2]. A strict interpretation of H and its probability of less than 0.01 does not, therefore, justify the conclusion that the means of the two chemical methods differ.

There is some reason to conjecture, however, that in practice the H test may be fairly insensitive to differences in variability, and so may be useful in the important "Behrens-Fisher problem" of comparing means without assuming equality of variances. Perhaps, for example, we could conclude that the means of the two chemical methods of Table 3.2 differ. The following considerations lend plausibility to this conjecture (and perhaps suggest extending it to other differences in form):

(*i*) The analysis of consistency referred to in Section 4.1 shows that if two symmetrical populations differ only by a scale factor about their common mean the H test is not consistent for small significance levels; in other words, below a certain level of significance there is no assurance that the null hypothesis of identical populations will be rejected, no matter how large the samples.

(*ii*) Consider the following extreme case: Samples of eight are drawn from two populations having the same mean but differing so much in variability that there is virtually no chance that any of the sample from the more variable population will lie within the range of the other sample. Furthermore, the median of the more variable population is at the common mean, so that its observations are as likely to lie above as to lie below the range of the sample from the less variable population. The actual distribution of H under these assumptions is easily computed from the binomial distribution with parameters 8 and $\frac{1}{2}$. Figure 4.1 shows the exact distribution of H under the null hypothesis that the two populations are completely identical, under the symmetrical alternative just described, and under a similar but skew alternative in which the probability is 0.65 that an observation from the more variable population will lie below the range of the other sample and 0.35 that it will lie above. Possible values of H under each hypothesis are those at which occur the risers in the corresponding step function of Figure 4.1, and the probabilities at these possible values of H are given by the tops of the risers. Figure 4.1 shows, for example, that samples in which seven observations from the more variable population lie above and one lies below the eight observations from the less variable population (so that the two values of R are 44 and 92, leading to an H of

6.353) would be judged by the H test to be significant at the 0.010 level using true probabilities (or at the 0.012 level using the χ^2 approximation), while such samples will occur about seven per cent of the time under the symmetrical alternative and about seventeen per cent under the other. In view of the extreme difference of the variances assumed in the alternatives, it seems rather striking that the cumulative distributions given in Figure 4.1 do not differ more than they do. At least in the case of the symmetrical alternative, the distribution for the null

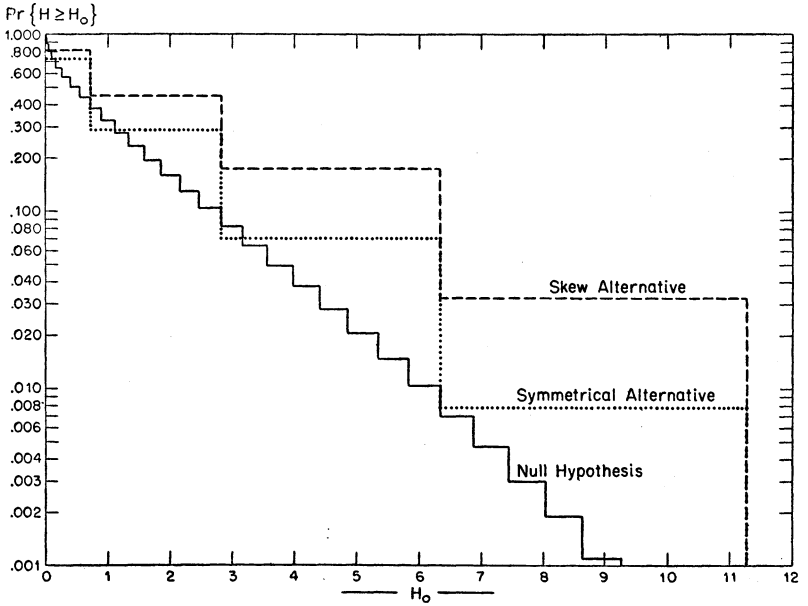

FIGURE 4.1. Distribution of H for two samples of 8, under the null hypothesis that the populations are identical and under two alternatives in which the means are the same but the variances are extremely different. (For further specification of the alternatives, see Section 4.2.)

hypothesis seems not too poor a partial smoothing, though on the whole it lies too low.

The applicability of the H test to the Behrens-Fisher problem, particularly in its two-tail form, merits further investigation.

5. RELATED TESTS

5.1. *Permutation Tests and Ranks*

The H test stems from two statistical methods, permutations of the data, and rank transformations.

Permutation tests, which to the best of our knowledge were first proposed by Fisher [8] in connection with a defense of the normality assumption, accept or reject the null hypothesis according to the probability of a test statistic among all relevant permutations of the observed numbers; a precise general formulation of the method is given by Scheffé [45]. Applications of the permutation method to important cases may be found in articles by Pitman [41, 42, 43] and by Welch [57].

The use of ranks—or more generally, of conventional numbers—instead of the observations themselves has been proposed often, and we do not know to whom this idea may be credited.[13] Its advantages have been summarized in Section 1.3. Its disadvantage is loss of information about exact magnitudes.

If in one-criterion variance analysis the permutation method based on the conventional F statistic is combined with the rank method, the result is the H test.

5.2. *Friedman's* χ_r^2

Two kinds of data must be distinguished in discussing tests for the equality of C population averages. The first kind consists of C independent random samples, one from each population. The second kind consists of C samples of equal size which are matched (that is, cross-classified or stratified, each stratum contributing one observation to each sample) according to some criterion which may affect the values of the observations. This distinction is, of course, exactly that between one-criterion variance analysis with equal sample sizes and two-criterion variance analysis with one observation per cell.

For comparing the weights of men and women, data of the first kind might be obtained by measuring a random sample of n_1 men and an independent random sample of n_2 women. Such data would ordinarily be analyzed by one-criterion variance analysis, as described in Section 1.2 above, which in the two-sample case is equivalent to the two-tail t test with n_1+n_2-2 degrees of freedom. The H test, or the two-sample version of it given by (3.4), would also be applicable.

Data of the second kind for the same problem might be obtained by selecting n ages (not necessarily all different) and for each age selecting at random one man and one woman. Such data would ordinarily be

[13] Our attention has been directed by Harold Hotelling to the use of ranks by Galton [12, Chaps. 4 and 5] in 1889. Churchill Eisenhart and I. Richard Savage have referred us to the extensive analyses of ranks by eighteenth century French mathematicians in connection with preference-ordering problems, specifically elections. The earliest work they mention is by Borda [1] in 1770, and they mention also Laplace [26] in 1778, Condorcet [3] in 1786, and Todhunter's summary of these and related writings [51, Secs. 690, 806, 989, 990]. Systematic treatment of ranks as a nonparametric statistical device, however, seems to commence with the work of Hotelling and Pabst [19] in 1936.

analyzed by two-criterion variance analysis, the between-sexes component being the one tested. This test would be equivalent to the two-tail t test of the mean difference, with $n-1$ degrees of freedom. Friedman's χ_r^2 [10], or the two-tail sign test which is its two-sample version, would be appropriate.[14]

The H test thus provides a rank test for data of the first kind, just as the χ_r^2 test does for data of the second kind. H makes it possible to test by ranks the significance of a grouping according to a single criterion. The effect of one criterion cannot be tested by χ_r^2 unless the observations in the different groups are matched according to a second criterion. On the other hand, if the data are matched H is not appropriate and χ_r^2 should be used.

5.3. *Wilcoxon's Two-Sample Test*

The H test in its general form is new, as far as we know,[15] but not its two-sample form.

5.3.1. *Wilcoxon (1945, 1947).* Wilcoxon was the first, we believe, to introduce the two-sample form. His first paper [61] considers the case of two samples of equal size and gives true probabilities for values of the smaller sum of ranks in the neighborhood of the 0.01, 0.02, and 0.05 probability levels for sample sizes from 5 to 10. A method of calculating the true probabilities is given. An example uses the mean-rank method for ties, interpreting the result in terms of a table for the no-ties situation.

In a second paper [62] on the case of two equal samples, Wilcoxon gives a normal approximation to the exact distribution, basing it on the theory of sampling without replacement from a finite uniform population, along the lines of Section 3.1 of the present paper. A table of 5 per cent, 2 per cent, and 1 per cent significance levels for the smaller total is given, covering sample sizes from 5 to 20.

5.3.2. *Festinger (1946).*[16] Wilcoxon's test was discovered independently by Festinger [7], who considers the case where the two sample sizes, n and m, are not necessarily equal. He gives a method of calculating true probabilities, and a table of two-tail 5 per cent and 1 per cent

[14] For other discussions of χ_r^2, see Kendall and Smith [21], Friedman [11], and Wallis [55].

[15] After an abstract [24] of a theoretical version [25] of the present paper was published we learned from T. J. Terpstra that similar work has been done at the Mathematical Center, Amsterdam, and that papers closely related to the H test will be published soon by himself [50] and by P. G. Rijkoort [44]; also that P. van Elteren and A. Benard are doing some research related to χ_r^2. References [50] and [44] propose tests based upon statistics similar to, but not identical with, H.

Alan Stuart tells us that H. R. van der Vaart (University of Leiden) has been planning a generalization of the Wilcoxon test to several samples.

P. V. Krishna Iyer has announced [23] "a non-parametric method of testing k samples." This brief announcement is not intelligible to us, but it states that "full details will be published in the *Journal of the Indian Society of Agricultural Research.*"

[16] We are indebted to Alan Stuart for calling our attention to Festinger's paper.

significance levels for n from 2 to 12 with m from n to $40-n$, and for n from 13 to 15 with m from n to $30-n$; and more extensive tables are available from him. A large proportion of the entries in Festinger's table, especially at the 5 per cent level, seem to be slightly erroneous.[21]

5.3.3. *Mann and Whitney (1947).* Mann and Whitney [28] made an important advance in showing that Wilcoxon's test is consistent for the null hypothesis that the two populations are identical against the alternative that the cumulative distribution of one lies entirely above that of the other.[17] They discuss the test in terms of a statistic U which, as they point out, is equivalent to Wilcoxon's sum of ranks (our R). When all observations from both samples are arranged in order, they count for each observation in one sample, say the first, the number of observations in the second sample that precede it. The sum of these counts for the first sample is called U. It is related to R, the sum of the ranks for the first sample, by[18]

$$U = R - \frac{n(n+1)}{2}. \tag{5.1}$$

They give a table showing the one-tail probability to three decimals for each possible value of U, for all combinations of sample sizes in which the larger sample is from three to eight.[19]

Hemelrijk [16] has pointed out recently that U, and consequently R for the two-sample case, may be regarded as a special case of Kendall's coefficient of rank correlation [20].

5.3.4. *Haldane and Smith (1948).*[20] Haldane and Smith [14] developed the Wilcoxon test independently in connection with the problem of deciding whether the probability of a hereditary trait appearing in a particular member of a sibship depends on his birthrank. They propose a test based on the sum of the birth-ranks of those members of a sibship having the trait—i.e., our R—where N is the number in the sibship and n is the number having the trait. They develop an approximate distribution from the theory of sampling from an infinite, continuous, uniform population, and approximate this by the unit normal deviate given in

[17] Actually the test is consistent under more general conditions; see Section 5.3.6 (*iv*).

[18] Mann and Whitney's version of this formula is a trifle different because they relate the count in the first sample (our terminology) to the sum of ranks in the other sample.

[19] We have recomputed the Mann-Whitney table to additional decimals. It agrees entirely with our computations.

[20] We are indebted to Alan Stuart for calling our attention to the Haldane and Smith paper.

Blair M. Bennett, University of Washington, is computing power functions for the Wilcoxon test against alternatives appropriate to the birth-order problem. Bennett emphasizes, in a personal communication, that the distribution of R under the null hypothesis corresponds to a partition problem which has been studied in the theory of numbers for centuries—in particular by Euler [6a, Chap. 16], who in 1748 considered closely related partition problems and their generating functions, and by Cauchy [2a, Numbers 225, 226]. In fact, Euler [6a, p. 252*] gives a table which is in part equivalent to that of Mann and Whitney [28]. This number-theoretic approach is discussed by Wilcoxon [61].

this paper as (3.4)—including the continuity adjustment, which they seem to be the first to use. They tabulate the means and variances of $6R$ for values of N from 2 to 20, with n from 1 to N. They also give a table of exact probabilities (not cumulated) for all possible values of n up to $N=12$.

Haldane and Smith discuss the problem of ties in connection with multiple births. They propose to assign to each member of each birth the rank of that birth. In our terminology, they give each member of a tied group the lowest of the ranks tied for, and give the next individual or group the next rank, not the rank after the highest in the group tied for. For a test in this case, they refer to the theory of sampling without replacement from a finite but non-uniform population.

With the Haldane-Smith method of handling ties, the difference between the ranks of two non-tied observations is one more than the number of distinct values or *groups* intervening between the two, regardless of the number of intervening individuals; with the mean-rank method, the difference is one more than the number of *observations* intervening, plus half the number of other observations having the same rank as either of the two observations being compared. The mean-rank method seems preferable when the cause of ties is measurement limitations on an effectively continuous variable, the Haldane-Smith method when the cause is actual identity. Unfortunately, the Haldane-Smith method does not lend itself so readily as does the mean-rank method to simple adjustment of the formulas for the no-ties case, since the necessary adjustments depend upon the particular ranks tied for, not merely the number of ties.

5.3.5. *White (1952).* Tables of critical values of R at two-tail significance levels of 5, 1, and 0.1 per cent for all sample sizes in which $N \leqq 30$ are given by White [59].[21] He suggests that ties be handled by the mean-rank method, not allowing for its effect on the significance level, or else by assigning the ranks so as to maximize the final probability, which may then be regarded as an upper limit for the true probability.

5.3.6. *Power of Wilcoxon's test.* The power of nonparametric tests in general, and of the H test in particular, is difficult to investigate; but

[21] Comparison of the 5 and 1 per cent levels given by White with Festinger's earlier and more extensive table [7] shows 104 disagreements among 392 comparable entries (78 disagreements among 196 comparisons at the 5 per cent level, and 26 among 196 at 1 per cent). In each disagreement, Festinger gives a lower critical value of the statistic, although both writers state that they have tabulated the smallest value of the statistic whose probability does not exceed the specified significance level. Three of the disagreements can be checked with the Mann-Whitney table [28]; in all three, White's entry agrees with Mann-Whitney's. In one additional case (sample sizes 4 and 11 at the 1 per cent level) we have made our own calculation and found Festinger's entry to have a true probability (0.0103) exceeding the stated significance level. The disagreements undoubtedly result from the fact that the distributions are discontinuous, so that exact 5 and 1 per cent levels cannot ordinarily be attained.

for the special case of Wilcoxon's two-sample test certain details have been discovered. Some that are interesting from a practical viewpoint are indicated below, but without the technical qualifications to which they are subject:

(*i*) Lehmann [27] has shown that the one-tail test is unbiased—that is, less likely to reject when the null hypothesis is true than when any alternative is true—but van der Vaart [52] has shown that the corresponding two-tail test may be biased.

(*ii*) Lehmann [27] has shown, on the basis of a theorem of Hoeffding's [17], that under reasonable alternative hypotheses, as under the null hypothesis, the distribution of $\sqrt{H}$ is asymptotically normal.

(*iii*). Mood [33] has shown that the asymptotic efficiency of Wilcoxon's test compared with Student's test, when both populations are normal with equal variance, is $3/\pi$, i.e., 0.955. Roughly, this means that $3/\pi$ is the limiting ratio of sample sizes necessary for the two tests to attain a fixed power. This result was given in lecture notes by E. J. G. Pitman at Columbia University in 1948; it was also given by van der Vaart [52]. To the best of our knowledge, Mood's proof is the first complete one.

(*iv*) Lehmann [27] and van Dantzig [15, 51a], generalizing the findings of Mann and Whitney [28], have shown that the test is consistent[12] if the probability differs from one-half that an observation from the first population will exceed one drawn independently from the second population (for one-tail tests the condition is that the probability differ from one-half in a stated direction). In addition van Dantzig [51a] gives inequalities for the power. The C-sample condition for consistency given by Kruskal (see Section 4.1) is a direct extension of the two sample condition given by Lehmann and van Dantzig.

5.4. *Whitney's Three-Sample Test*

Whitney [60] has proposed two extensions of the Wilcoxon test to the three-sample case. Neither of his extensions, which are expressed in terms of inversions of order rather than in terms of ranks, is equivalent to our H test for $C=3$, since Whitney seeks tests with power against more specific alternatives than those appropriate to the H test.

Whitney arrays all three samples in a single ranking and then defines U as the number of times in which an observation from the second sample precedes an observation from the first and V as the number of times in which an observation from the third sample precedes one from the first.[22]

[22] U and V are not determined by R_1, R_2, and R_3, nor vice versa, though

$$U + V = R_1 - \tfrac{1}{2}n_1(n_1 + 1)$$

Whitney's first test, which rejects the null hypothesis of equality of the populations if both U and V are too small (alternatively, too large), is suggested when the alternative is that the cumulative distribution of the first population lies above (alternatively, below) those of both the second and third populations. His second test, which rejects if U is too large and V is too small, is suggested when the alternative is that the cumulative distribution of the first population lies below that of the second and above that of the third.

5.5 *Terpstra's C-sample Test.*

Terpstra [50a] has proposed and investigated a test appropriate for alternatives similar to those of Whitney's second test, but extending to any number of populations.

5.6. *Mosteller's C-Sample Test*

Mosteller [34] has proposed a multi-decision procedure for accepting either the null hypothesis to which the H test is appropriate or one of the C alternatives that the ith population is translated to the right (or left) of the others. His criterion is the number of observations in the sample containing the largest observation that exceed all observations in other samples. This procedure has been discussed further by Mosteller and Tukey [35].

5.7. *Fisher and Yates' Normalized Ranks*

Fisher and Yates have proposed [9, Table XX] that each observation be replaced not by its simple rank but by a normalized rank, defined as the average value of the observation having the corresponding rank in samples of N from a normal population with mean of zero and standard deviation of one. They propose that ordinary one-criterion variance analysis then be applied to these normalized ranks. Ehrenberg [5] has suggested as a modification using the values of a random sample of N from the standardized normal population.

Two advantages might conceivably be gained by replacing the observations by normalized ranks or by some other set of numbers instead of by simple ranks. First, it might be that the distribution theory would be simplified. Quite a large class of such transformations, for example, lead to tests whose distribution is asymptotically $\chi^2(C-1)$; but for some transformations the χ^2 approximation may be satisfactory at smaller sample sizes than for others, thus diminishing the area of need for special tables and approximations such as those presented in Sec. 6.

Second, the power of the test might be greater against important classes of alternatives.

Whether either of these possible advantages over ranks is actually realized by normalized ranks, or by any other specific transformation, has not to our knowledge been investigated. Offhand, it seems intuitively plausible that the χ^2 distribution might be approached more rapidly with normalized ranks, or some other set of numbers which resemble the normal form more than do ranks. On the other hand, it seems likely that if there is such an advantage it is not very large, partly because the distribution of means from a uniform population approaches normality rapidly as sample size increases, and partly because (as Section 6 indicates) the distribution of H approaches the χ^2 distribution quite rapidly as sample sizes increase. As to power, we have no suggestions, except the obvious one that the answer is likely to differ for different alternatives of practcal interest.[23]

5.8. *Other Related Tests*

A number of tests have been proposed which have more or less the same purpose as H and are likewise non-parametric. We mention here only two of the principal classes of these.

5.8.1. *Runs.* Wald and Wolfowitz [53] have proposed for the two-sample case that all observations in both samples be arranged in order of magnitude, that the observations then be replaced by designations A or B, according to which sample they represent, and that the number of runs (i.e., groups of consecutive A's or consecutive B's) be used to test the null hypothesis that both samples are from the same population. The distribution theory of this test has been discussed by Stevens [48], Wald and Wolfowitz [53], Mood [31], Krishna Iyer [22], and others; and Swed and Eisenhart [49] have provided tables covering all cases in which neither sample exceeds 20. For larger samples, normal approximations are given by all the writers mentioned. Wald and Wolfowitz discussed the consistency of the test, and later Wolfowitz [63] discussed its asymptotic power. An extension to cases of three or more samples has been given by Wallis [56], based on the distribution theory of Mood and Krishna Iyer.

5.8.2. *Order statistics.* Westenberg [58] has suggested a test for the two-sample case utilizing the number of observations in each sample above the median of the combined samples. Mood and Brown [32, pp.

[23] When the true distributions are normal, Hoeffding [18] has shown that in many cases, including at least some analysis of variance ones, the test based on normalized ranks becomes as powerful as that based on the actual observations, when the sample sizes increase toward infinity.

394–5, 398–9] have discussed the test further and generalized it to several samples. Massey [29] has generalized the test further by using other order statistics of the combined samples as a basis for a more refined classification.

6. SIGNIFICANCE LEVELS, TRUE AND APPROXIMATE

6.1. *True Significance Levels*

6.1.1. *Two samples.* Festinger [7], Haldane and Smith [14], Mann and Whitney [28], White [59], and Wilcoxon [61, 62] have published tables for the two-sample case. These are described in Section 5.3. They are exact only if ties are absent or are handled by the random-rank method, but our guess is that they will also serve well enough if the mean-rank method is used and there are not too many ties.

6.1.2. *Three samples.* (*i*) *Five or fewer observations in each sample.* For each of these cases, Table 6.1 shows three pairs of values of H and their probabilities of being equalled or exceeded if the null hypothesis is true[24]. Each pair brackets as closely as possible the 10, 5, or 1 per cent level, except that in some cases one or both members of a pair are missing because H can take only a small number of values. The final sentence of Section 6.1.1, about ties, applies to Table 6.1 also.

(*ii*) *More than five observations in each sample.* No exact tables are available for these cases. Our recommendation is that the χ^2 approximation be used. Only at very small significance levels (less than 1 per cent, say) and sample sizes only slightly above five is there likely to be appreciable advantage to the more complicated Γ and B approximations described in Section 6.2. This recommendation is based only on the comparisons shown in Table 6.1, no true probabilities having been computed in this category.

(*iii*) *Intermediate cases.* No exact tables are available here. The Γ and B approximations probably should be resorted to if more than roughly approximate probabilities are required. Except at very low significance levels or with very small samples, the Γ approximation, which is simpler, should serve. This recommendation is not very firm, however, since we have computed no true probabilities in this category.

6.1.3. *More than three samples.* Since we have computed no true probabilities for more than three samples, our recommendations here

[24] These computations and others used for this paper were made by John P. Gilbert with the assistance of Billy L. Foster, Thomas O. King, and Roland Silver. Space prevents reproducing all or even most of the results, but we hope to file them in such a way that interested workers may have access to them. We have the true joint distributions of R_1, R_2, and R_3 under the null hypothesis for n_1, n_2, and n_3, each from 1 through 5, and the true distribution of H under the same conditions, except that for some cases we have probabilities only for those values of H exceeding the upper twenty per cent level.

must be entirely tentative. It seems safe to use the χ^2 approximation when all samples are as large as five. If any sample is much smaller than five, the Γ or B approximation should probably be used, especially at low significance levels, though the importance of this presumably is less the larger the proportion of samples of more than five.

6.2. *Approximate Significance Levels*

6.2.1. χ^2 *approximation.* This is the approximation discussed in Sections 1, 2, and 3. The most extensive single table is that of Hald and Sinkbaek [13], though the table in almost any modern statistics text will ordinarily suffice.

6.2.2. Γ *approximation.* This utilizes the incomplete-Γ distribution by matching the variance as well as the true mean of H. The mean, or expected value, of H under the null hypothesis is [25]

$$E = C - 1 \tag{6.1}$$

and the variance is

$$V = 2(C-1) - \frac{2[3C^2 - 6C + N(2C^2 - 6C + 1)]}{5N(N+1)} - \frac{6}{5}\sum_{i=1}^{C}\frac{1}{n_i}. \tag{6.2}$$

One way of applying the approximation is to enter an ordinary χ^2 table taking $\chi^2 = 2HE/V$ and degrees of freedom $f = 2E^2/V$. Note that the degrees of freedom will not ordinarily be an integer, so interpolation will be required in both χ^2 and the degrees of freedom if the four bounding tabular entries do not define the probability accurately enough.[25]

6.2.3. B *approximation.* This utilizes the incomplete-B distribution by matching the true maximum as well as the mean and variance of H. The maximum value of H is [25]

$$M = \frac{N^3 - \sum_{i=1}^{C} n_i^3}{N(N+1)}. \tag{6.3}$$

To apply the approximation, K. Pearson's table of the incomplete-B distribution [39] may be employed, but it is usually more convenient to use the F distribution, a form of the incomplete-B distribution, since

[25] The Γ approximations shown in Table 6.1 were based on K. Pearson's table of the incomplete-Γ function [40]. In Pearson's notation, the required probability is $1 - I(u, p)$, where $u = \mathrm{H}/\sqrt{V}$ and $p = E^2/V - 1$. We used linear double interpolation, which on a few tests seemed to be satisfactory in the region of interest.

tables of F are widely accessible to statisticians.[26] We set

$$F = \frac{H(M - E)}{E(M - H)} \tag{6.4}$$

with degrees of freedom (not usually integers)

$$f_1 = E \cdot \frac{E(M - E) - V}{\frac{1}{2}MV}, \tag{6.5}$$

$$f_2 = (M - E) \cdot \frac{E(M - E) - V}{\frac{1}{2}MV} = \frac{M - E}{E} \cdot f_1. \tag{6.6}$$

The probability may then be obtained by three-way interpolation in the F tables or by using Paulson's approximation [36], according to which the required probability, P, is the probability that a unit normal deviate will exceed

$$K_P = \frac{(1 - 2/9f_2)F' + 2/9f_1 - 1}{\sqrt{2F'^2/9f_2 + 2/9f_1}} \tag{6.7}$$

where $F' = \sqrt[3]{F}$.

As an illustration, suppose $C=3$, $n_1=5$, $n_2=4$, $n_3=3$, and $H=5.6308$. From (6.1), (6.2), and (6.3) we find $E=2$, $V=3.0062$, and $M=9.6923$. Substituting these into (6.4), (6.5), and (6.6) gives $F=5.332$, $f_1=1.699$, and $f_2=6.536$. Then (6.7) gives $K_P=1.690$, for which the normal distribution shows a probability of 0.046. This may be compared with the true probability of 0.050, the χ^2 approximation of 0.060, and the Γ approximation of 0.044, shown in Table 6.1.[27]

---

[26] The most detailed table of the F distribution is that of Merrington and Thompson [30].

[27] The B approximations shown in Table 6.1 are based on K. Pearson's table of the incomplete-B function [39]. In Pearson's notation, the required probability is $1 - I_x(p, q)$, where $x = H/M$, $p = \frac{1}{2}f_1$, and $q = \frac{1}{2}f_2$. To simplify the three-way interpolation, the following device (based on the relation of the incomplete-B to the binomial distribution, and of the binomial to the normal distribution) was used: *First*, let p_0, q_0, and x_0 be the tabulated arguments closest to p, q, and x, and as a first approximation to the required probability take $1 - I_{x_0}(p_0, q_0)$. *Second*, add to this first approximation the probability that a unit normal deviate will not exceed (in algebraic, not absolute, value)

$$K = \frac{p - \frac{1}{2} - x(p + q - 1)}{\sqrt{x(1 - x)(p + q - 1)}}$$

Third, subtract from this the probability that a unit normal deviate will not exceed K_0, where K_0 is defined like K but in terms of p_0, q_0, and x_0. This method of interpolation was compared at three points with the trivariate Everett formula to third differences as presented by Pearson [39, Introduction]. The results were not excellent, but seemed to suffice for the present purposes.

Strictly speaking, all our statements and numerical results concerning the B approximation (including entries in Table 6.1) actually apply to that approximation based on Pearson's tables in combination with this normal interpolation device.

Values calculated in this way will not in general agree precisely with those calculated by interpolating in the F tables or by using Paulson's approximation, though the example in the text agrees to three decimals.

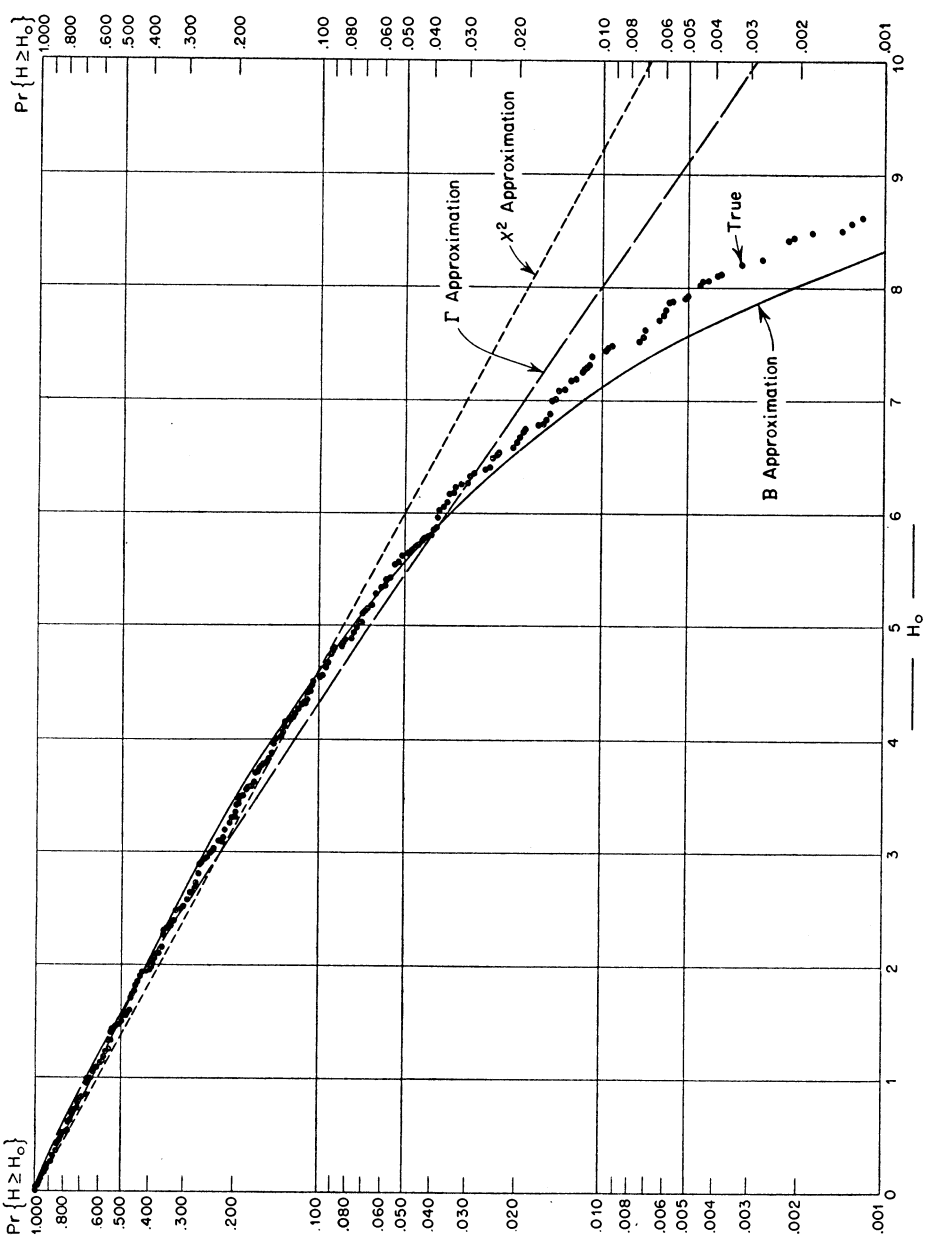

FIGURE 6.1. True distribution (under the null hypothesis) of H for three samples of sizes 5, 4, and 3, and the χ^2, Γ, and B approximations.

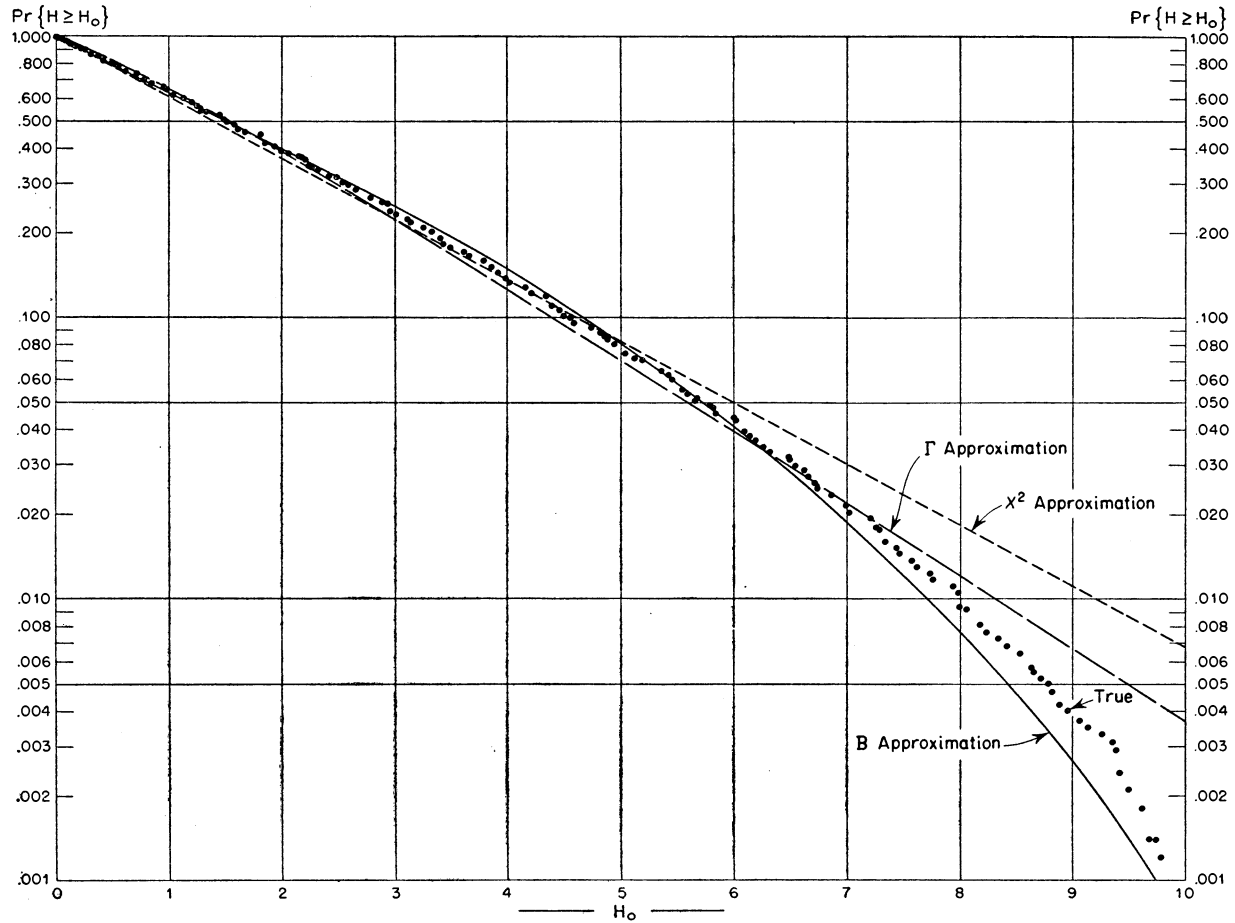

FIGURE 6.2. True distribution (under the null hypothesis) of H for three samples each of size 5, and the χ^2, Γ, and B approximations.

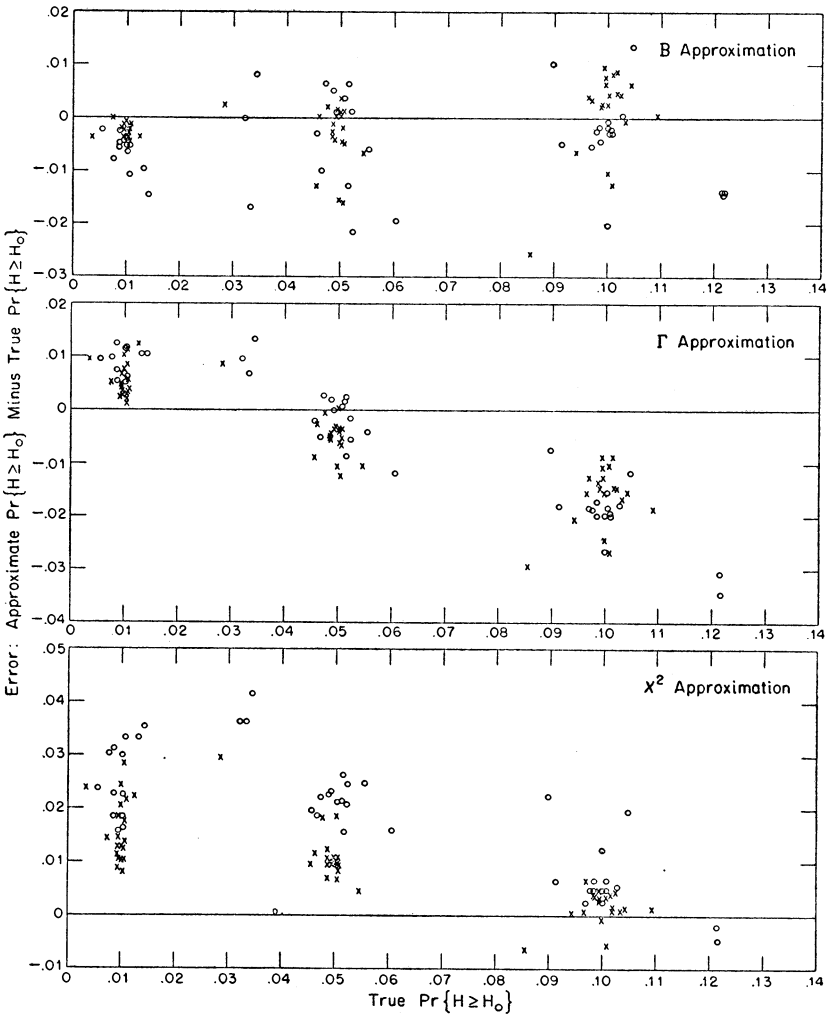

FIGURE 6.3. Comparison of the true and approximate significance probabilities for H in the neighborhoods of the 1, 5, and 10 per cent points, for three samples of sizes 2 to 5. Crosses indicate that the smallest sample size exceeds 2, circles that it is 2. Cases involving samples of 1 and a few involving samples of 2 have been omitted.

TABLE 6.1

TRUE DISTRIBUTION OF H FOR THREE SAMPLES, EACH OF SIZE FIVE OR LESS, IN THE NEIGHBORHOOD OF THE 10, 5, AND 1 PER CENT POINTS; AND COMPARISON WITH THREE APPROXIMATIONS

The probabilities shown are the probabilities under the null hypothesis that H will equal or exceed the values in the column headed "H"

Sample Sizes			H	True Probability	Approximate minus true probability		
n_1	n_2	n_3			χ^2	Γ (Linear Interp.)	B (Normal Interp.)
2	1	1	2.7000	.500	−.241	−.309	−.500
2	2	1	3.6000	.267	−.101	−.167	−.267
2	2	2	4.5714	.067	+.035	−.007	−.067
			3.7143	.200	−.044	−.083	+.010
3	1	1	3.2000	.300	−.098	−.180	−.300
3	2	1	4.2857	.100	+.017	−.040	−.100
			3.8571	.133	+.012	−.045	−.042
3	2	2	5.3572	.029	+.040	+.083	−.029
			4.7143	.048	+.047	+.012	+.014
			4.5000	.067	+.039	+.003	+.020
			4.4643	.105	+.002	−.033	−.014
3	3	1	5.1429	.043	+.034	−.010	−.043
			4.5714	.100	+.002	−.046	−.062
			4.0000	.129	+.007	−.041	−.024
3	3	2	6.2500	.011	+.033	+.012	−.011
			5.3611	.032	+.036	+.010	+.001
			5.1389	.061	+.016	−.012	−.019
			4.5556	.100	+.002	−.027	−.020
			4.2500	.121	−.002	−.031	−.014
3	3	3	7.2000	.004	+.024	+.010	−.004
			6.4889	.011	+.028	+.011	−.001
			5.6889	.029	+.030	+.009	+.003
			5.6000	.050	+.011	−.010	−.015
			5.0667	.086	−.006	−.029	−.026
			4.6222	.100	−.001	−.025	−.010
4	1	1	3.5714	.200	−.032	−.114	−.200
4	2	1	4.8214	.057	+.033	−.017	−.057
			4.5000	.076	+.029	−.022	−.047
			4.0179	.114	+.020	−.032	−.056
4	2	2	6.0000	.014	+.036	+.010	−.014
			5.3333	.033	+.036	+.007	−.017
			5.1250	.052	+.025	−.006	−.021
			4.3750	.100	+.012	−.020	−.002
			4.1667	.105	+.020	−.012	+.014

TABLE 6.1 (Continued)

Sample Sizes			H	True Probability	Approximate minus true probability		
n_1	n_2	n_3			χ^2	Γ (Linear Interp.)	B (Normal Interp.)
4	3	1	5.8333	.021	+.033	−.001	−.021
			5.2083	.050	+.024	−.016	−.037
			5.0000	.057	+.025	−.016	−.034
			4.0556	.093	+.039	−.005	+.014
			3.8889	.129	+.014	−.028	−.003
4	3	2	6.4444	.009	+.031	+.012	−.002
			6.4222	.010	+.030	+.011	−.004
			5.4444	.047	+.019	−.005	−.010
			5.4000	.052	+.016	−.008	−.013
			4.5111	.098	+.006	−.020	−.004
			4.4667	.101	+.006	−.020	−.003
4	3	3	6.7455	.010	+.024	+.010	−.001
			6.7091	.013	+.022	+.007	−.003
			5.7909	.046	+.010	−.009	−.013
			5.7273	.050	+.007	−.012	−.015
			4.7091	.094	+.001	−.021	−.006
			4.7000	.101	−.006	−.027	−.012
4	4	1	6.6667	.010	+.026	+.002	−.010
			6.1667	.022	+.024	−.005	−.020
			4.9667	.048	+.036	−.003	−.009
			4.8667	.054	+.034	−.005	−.009
			4.1667	.082	+.042	+.002	+.016
			4.0667	.102	+.029	−.011	+.007
4	4	2	7.0364	.006	+.024	+.010	−.002
			6.8727	.011	+.021	+.006	−.005
			5.4545	.046	+.020	−.002	−.003
			5.2364	.052	+.021	−.002	+.001
			4.5545	.098	+.005	−.019	−.003
			4.4455	.103	+.006	−.018	+.000
4	4	3	7.1439	.010	+.018	+.007	−.002
			7.1364	.011	+.018	+.006	−.003
			5.5985	.049	+.012	−.005	−.004
			5.5758	.051	+.011	−.006	−.005
			4.5455	.099	+.004	−.015	+.003
			4.4773	.102	+.004	−.014	+.004
4	4	4	7.6538	.008	+.014	+.005	.000
			7.5385	.011	+.012	+.003	−.002
			5.6923	.049	+.009	−.006	−.002
			5.6538	.054	+.005	−.010	−.007
			4.6539	.097	+.001	−.015	+.004
			4.5001	.104	+.001	−.015	+.007
5	1	1	3.8571	.143	+.003	−.109	−.143

TABLE 6.1 (Continued)

Sample Sizes			H	True Probability	Approximate minus true probability		
n_1	n_2	n_3			χ^2	Γ (Linear Interp.)	B (Normal Interp.)
5	2	1	5.2500	.036	+.037	−.006	−.036
			5.0000	.048	+.034	+.011	−.037
			4.4500	.071	+.037	−.012	−.020
			4.2000	.095	+.027	−.022	−.018
			4.0500	.119	+.013	−.036	−.024
5	2	2	6.5333	.008	+.030	+.010	−.008
			6.1333	.013	+.033	+.010	−.010
			5.1600	.034	+.041	+.013	+.008
			5.0400	.056	+.025	−.004	−.006
			4.3733	.090	+.022	−.007	+.010
			4.2933	.122	−.005	−.034	−.014
5	3	1	6.4000	.012	+.029	+.002	−.012
			4.9600	.048	+.036	−.004	−.010
			4.8711	.052	+.036	−.004	−.009
			4.0178	.095	+.039	−.002	+.018
			3.8400	.123	+.024	−.016	+.010
5	3	2	6.9091	.009	+.023	+.007	−.006
			6.8218	.010	+.023	+.007	−.006
			5.2509	.049	+.023	−.000	+.001
			5.1055	.052	+.026	+.003	+.006
			4.6509	.091	+.006	−.018	−.005
			4.4945	.101	+.005	−.020	−.003
5	3	3	6.9818	.010	+.020	+.008	−.002
			6.8606	.011	+.022	+.008	−.001
			5.4424	.048	+.018	−.000	+.002
			5.3455	.050	+.019	+.000	+.004
			4.5333	.097	+.007	−.013	+.004
			4.4121	.109	+.001	−.018	+.000
5	4	1	6.9545	.008	+.023	+.002	−.008
			6.8400	.011	+.022	−.000	−.011
			4.9855	.044	+.038	+.002	−.001
			4.8600	.056	+.032	−.005	−.005
			3.9873	.098	+.038	+.001	+.018
			3.9600	.102	+.036	−.000	+.018
5	4	2	7.2045	.009	+.018	+.005	−.005
			7.1182	.010	+.018	+.005	−.005
			5.2727	.049	+.023	+.002	+.005
			5.2682	.050	+.021	+.000	+.004
			4.5409	.098	+.005	−.017	−.002
			4.5182	.101	+.004	−.018	−.002

TABLE 6.1 (Continued)

Sample Sizes			H	True Probability	Approximate minus true probability		
n_1	n_2	n_3			χ^2	Γ (Linear Interp.)	B (Normal Interp.)
5	4	3	7.4449	.010	+.014	+.004	−.004
			7.3949	.011	+.014	+.004	−.004
			5.6564	.049	+.010	−.005	−.004
			5.6308	.050	+.010	−.006	−.004
			4.5487	.099	+.004	−.013	+.003
			4.5231	.103	+.001	−.016	−.000
5	4	4	7.7604	.009	+.011	+.003	−.002
			7.7440	.011	+.010	+.002	−.003
			5.6571	.049	+.010	−.004	+.000
			5.6176	.050	+.010	−.004	+.001
			4.6187	.100	−.001	−.016	+.003
			4.5527	.102	+.001	−.014	+.005
5	5	1	7.3091	.009	+.016	−.002	−.009
			6.8364	.011	+.022	+.001	−.009
			5.1273	.046	+.031	−.003	−.005
			4.9091	.053	+.032	−.002	−.002
			4.1091	.086	+.042	+.007	+.020
			4.0364	.105	+.028	−.007	+.008
5	5	2	7.3385	.010	+.016	+.004	−.004
			7.2692	.010	+.016	+.004	−.004
			5.3385	.047	+.022	+.003	+.006
			5.2462	.051	+.022	+.002	+.007
			4.6231	.097	+.002	−.018	−.005
			4.5077	.100	+.005	−.016	−.001
5	5	3	7.5780	.010	+.013	+.004	−.001
			7.5429	.010	+.013	+.004	−.002
			5.7055	.046	+.012	−.003	+.000
			5.6264	.051	+.009	−.005	−.002
			4.5451	.100	+.003	−.012	+.007
			4.5363	.102	+.002	−.014	+.005
5	5	4	7.8229	.010	+.010	+.003	−.002
			7.7914	.010	+.010	+.003	−.002
			5.6657	.049	+.010	−.003	+.001
			5.6429	.050	+.009	−.003	+.001
			4.5229	.099	+.005	−.009	+.010
			4.5200	.101	+.004	−.010	+.008
5	5	5	8.0000	.009	+.009	+.003	−.002
			7.9800	.010	+.008	+.002	−.003
			5.7800	.049	+.007	−.005	−.001
			5.6600	.051	+.008	−.004	+.001
			4.5600	.100	+.003	−.010	+.008
			4.5000	.102	+.004	−.009	+.009

6.3. *Comparisons of True and Approximate Significance Levels*

Figures 6.1 and 6.2 show the true probabilities and the χ^2, Γ, and B approximations when the sample sizes are 3, 4, and 5, and when they are all 5.[28]

For each entry in Table 6.1 the probabilities given by the three approximations have been computed and their errors recorded in the last three columns of the table. In Figure 6.3 these errors are graphed against the true probabilities. To avoid confusing this figure, sample sizes have not been indicated; cases involving samples of one have been omitted, and cases involving samples of two have been distinguished from those in which the smallest sample exceeds two.

7. REFERENCES

[1] Borda, Jean Charles, "Mémoire sur les élections au scrutin," *Mémoires de l'Académie royale des Sciences de Paris pour l'Année 1781*, pp. 657–65.

[2] Brownlee, K. A., *Industrial Experimentation*, Third American Edition, Brooklyn, Chemical Publishing Company, 1949.

[2a] Cauchy, D'Augustin, "Oeuvres complètes," Series 1, Volume 8, Paris, Gauthier-Villars et Fils, 1893.

[3] Condorcet, le Marquis de (Marie Jean Antoine Nicolas Caritat), *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris, 1785, pp. lvii, clxxvii ff.

[4] DuBois, Philip, "Formulas and tables for rank correlation," *Psychological Record*, 3 (1939), 46–56.

[5] Ehrenberg, A. S. C., "Note on normal transformations of ranks," *British Journal of Psychology, Statistical Section*, 4 (1951), 133–4.

[6] Eudey, M. W., *On the treatment of discontinuous random variables*, Technical Report Number 13, Statistical Laboratory, University of California (Berkeley), 1949.

[6a] Euler, Leonhard, "Introduction à l'analyse infinitésimale," (translated from the Latin edition of 1748 into French by J. B. Labey), Vol. 1, Paris, Chez Barrois, 1796.

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

[8] Fisher, R. A., *The Design of Experiments*, Edinburgh, Oliver and Boyd Ltd., 1935 and later.

[9] Fisher, Ronald A., and Yates, Frank, *Statistical Tables for Biological, Agricultural and Medical Research*, Edinburgh, Oliver and Boyd Ltd., 1938 and later.

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

[11] Friedman, Milton, "A comparison of alternative tests of significance for the problem of m rankings," *Annals of Mathematical Statistics*, 11 (1940), 86–92.

[28] All four figures in this paper are the work of H. Irving Forman.

[12] Galton, Sir Francis, *Natural Inheritance*, London, Macmillan and Co., 1889.
[13] Hald, A., and Sinkbaek, S. A., "A table of percentage points of the χ^2-distribution," *Skandinavisk Aktuarietidskrift*, 33 (1950), 168–75.
[14] Haldane, J. B. S., and Smith, Cedric A. B., "A simple exact test for birth-order effect," *Annals of Eugenics*, 14 (1947–49), 117–24.
[15] Hemelrijk, J., "A family of parameterfree tests for symmetry with respect to a given point. II," *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen*, 53 (1950), 1186–98.
[16] Hemelrijk, J., "Note on Wilcoxon's two-sample test when ties are present," *Annals of Mathematical Statistics*, 23 (1952), 133–5.
[17] Hoeffding, Wassily, "A class of statistics with asymptotically normal distributions," *Annals of Mathematical Statistics*, 19 (1948), 293–325.
[18] Hoeffding, Wassily, "Some powerful rank order tests" (abstract), *Annals of Mathematical Statistics*, 23 (1952), 303.
[18a] Horn, Daniel, "A correction for the effect of tied ranks on the value of the rank difference correlation coefficient," *Journal of Educational Psychology*, 33 (1942), 686–90.
[19] Hotelling, Harold, and Pabst, Margaret Richards, "Rank correlation and tests of significance involving no assumption of normality," *Annals of Mathematical Statistics*, 7 (1936), 29–43.
[20] Kendall, Maurice G., *Rank Correlation Methods*, London, Charles Griffin and Company, 1948.
[21] Kendall, Maurice G., and Smith, B. Babington, "The problem of m rankings," *Annals of Mathematical Statistics*, 10 (1939), 275–87.
[22] Krishna Iyer, P. V., "The theory of probability distributions of points on a line," *Journal of the Indian Society of Agricultural Statistics*, 1 (1948), 173–95.
[23] Krishna Iyer, P. V., "A non-parametric method of testing k samples," *Nature*, 167 (1951), 33.
[24] Kruskal, William H., "A nonparametric analogue based upon ranks of one-way analysis of variance" (abstract), *Annals of Mathematical Statistics*, 23 (1952), 140.
[25] Kruskal, William H., "A nonparametric test for the several sample problem," *Annals of Mathematical Statistics*, 23 (1952), 525–40.
[26] Laplace, Pierre Simon, *A Philosophical Essay on Probabilities*, New York, Dover Publications, Inc., 1951 (first edition 1814).
[27] Lehmann, E. L., "Consistency and unbiasedness of certain non-parametric tests," *Annals of Mathematical Statistics*, 22 (1951), 165–79.
[28] Mann, H. B., and Whitney, D. R., "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, 18 (1947), 50–60.
[29] Massey, Frank J., Jr., "A note on a two-sample test," *Annals of Mathematical Statistics*, 22 (1951), 304–6.
[30] Merrington, Maxine, and Thompson, Catherine M., "Tables of percentage points of the inverted Beta (F) distribution," *Biometrika*, 33 (1943), 73–88.
[31] Mood, A. M., "The distribution theory of runs," *Annals of Mathematical Statistics*, 11 (1940), 367–92.
[32] Mood, Alexander McFarlane, *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950.

[33] Mood, A. M. Unpublished manuscript, submitted to *Annals of Mathematical Statistics.*

[34] Mosteller, Frederick, "A k-sample slippage test for an extreme population," *Annals of Mathematical Statistics,* 19 (1948), 58–65.

[35] Mosteller, Frederick, and Tukey, John W., "Significance levels for a k-sample slippage test," *Annals of Mathematical Statistics,* 21 (1950), 120–3.

[36] Paulson, Edward, "An approximate normalization of the analysis of variance distribution," *Annals of Mathematical Statistics,* 13 (1942), 233–5.

[37] Pearson, E. S., "On questions raised by the combination of tests based on discontinuous distributions," *Biometrika,* 37 (1950), 383–98.

[38] Pearson, Karl, "On a certain double hypergeometrical series and its representation by continuous frequency surfaces," *Biometrika,* 16 (1924), 172–88.

[39] Pearson, Karl, editor, *Tables of the Incomplete Beta Function,* London, Biometrika Office, 1934.

[40] Pearson, Karl, editor, *Tables of the Incomplete* Γ*-Function,* London, Biometrika Office, 1951 (reissue).

[41] Pitman, E. J. G., "Significance tests which may be applied to samples from any populations," *Supplement to the Journal of the Royal Statistical Society,* 4 (1937), 119–30.

[42] Pitman, E. J. G., "Significance tests which may be applied to samples from any populations. II. The correlation coefficient test," *Supplement to the Journal of the Royal Statistical Society,* 4 (1937), 225–32.

[43] Pitman, E. J. G., "Significance tests which may be applied to samples from any populations. III. The analysis of variance test," *Biometrika,* 29 (1937), 322–35.

[44] Rijkoort, P. G., "A generalization of Wilcoxon's test," *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen,* 53 (1952).

[45] Scheffé, Henry, "Statistical inference in the non-parametric case," *Annals of Mathematical Statistics,* 14 (1943), 305–32.

[46] Snedecor, George W., *Statistical Methods,* Ames, Iowa State College Press, 1937 and later.

[47] Spława-Neyman, Jerzy, "Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych. (Sur les applications de la théorie des probabilités aux expériences agricoles. Essay des principes)," *Roczniki Nauk Rolniczych,* 10 (1923), 1–51. (Polish with German summary.)

[48] Stevens, W. L., "Distribution of groups in a sequence of alternatives," *Annals of Eugenics,* 9 (1939), 10–17.

[48a] 'Student,' "An experimental determination of the probable error of Dr. Spearman's correlation coefficient," *Biometrika,* 13 (1921), 263–82. Reprinted in *'Student's' Collected Papers* (edited by E. S. Pearson and John Wishart), London, Biometrika Office, n.d., 70–89.

[49] Swed, Frieda S., and Eisenhart, C., "Tables for testing randomness of grouping in a sequence of alternatives," *Annals of Mathematical Statistics,* 14 (1943), 66–87.

[50] Terpstra, T. J., "A non-parametric k-sample test and its connection with the H test." Unpublished manuscript.

[50a] Terpstra, T. J., "The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking," *Indagationes Mathematicae,* 14 (1952), 327–33.

[51] Todhunter, Isaac, *A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace*, New York, Chelsea Publishing Company, 1949 (first edition 1865).

[51a] van Dantzig, D., "On the consistency and the power of Wilcoxon's two sample test," *Indagationes Mathematicae*, 13 (1951), 1–8; also *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen*, 54 (1951), 1–8.

[52] van der Vaart, H. R., "Some remarks on the power of Wilcoxon's test for the problem of two samples," *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen*, 53 (1950), 494–506, 507–20.

[53] Wald, A., and Wolfowitz, J., "On a test whether two samples are from the same population," *Annals of Mathematical Statistics*, 11 (1940), 147–62.

[54] Wald, A., and Wolfowitz, J., "Statistical tests based on permutations of the observations," *Annals of Mathematical Statistics*, 15 (1944), 358–72.

[55] Wallis, W. Allen, "The correlation ratio for ranked data," *Journal of the American Statistical Association*, 34 (1939), 533–8.

[56] Wallis, W. Allen, "Rough-and-ready statistical tests," *Industrial Quality Control*, 8 (1952), 35–40.

[57] Welch, B. L., "On the z-test in randomized blocks and Latin Squares," *Biometrika*, 29 (1937), 21–52.

[58] Westenberg, J., "Significance test for median and interquartile range in samples from continuous populations of any form," *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen*, 51 (1948), 252–61.

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

[60] Whitney, D. R., "A bivariate extension of the U statistic," *Annals of Mathematical Statistics*, 22 (1951), 274–82.

[61] Wilcoxon, Frank, "Individual comparisons by ranking methods," *Biometrics Bulletin* (now *Biometrics*), 1 (1945), 80–3.

[62] Wilcoxon, Frank, "Probability tables for individual comparisons by ranking methods," *Biometrics*, 3 (1947), 119–22.

[63] Wolfowitz, J., "Non-parametric statistical inference," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (edited by Jerzy Neyman), Berkeley and Los Angeles, University of California Press, 1949, 93–113.