NON-PARAMETRIC STATISTICS FOR PSYCHOLOGICAL RESEARCH

LINCOLN E. MOSES
Teachers College, Columbia University

It has been said that "everybody believes in the law of errors [the normal distribution], the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."[1] There are excellent theoretical reasons to explain the preeminent position which the normal distribution has held in the development of statistical theory.[2]

On the other hand, at some time or other nearly every experimenter must work with samples which he knows do not come from a normal distribution. If he knows what the distribution actually is then he may find a transformation such that his transformed data *are* observations from a normal distribution, or he may find a special theory already worked out (as for, say, the Poisson distribution). More often he has no such knowledge of the population distribution, and then he must choose between applying the textbook methods in violation of their underlying assumptions, or of finding valid techniques which have no underlying assumptions concerning the shape of the parent population.

Until about fifteen years ago this was merely Hobson's choice, since about the only distribution-free methods were rank correlation and χ^2 tests. But there has recently been a great growth in statistical methodology which provides the experimenter with tools free of assumptions about the population distribution. These techniques are generally referred to as Non-Parametric Methods, or sometimes, Distribution Free Methods.

It is the purpose of this paper to present some of the principal methods, and an intuitive explanation of their rationale, properties, and applicability, with a view to facilitating their use by workers in psychological research.

In many, but not all, of the methods discussed, the data to which the tests are applied are not the original measurements in the sample

[1] Cramér cites Poincare's quotation of this famous remark by Lippman. (CRAMÉR, H., *Mathematical methods of statistics*, Princeton Univ. Press, 1946, p. 232.)

[2] Perhaps the most outstanding of these is the so-called Central Limit Theorem which specifies (roughly) that means of "large" (enough) samples from *any* population (except for some pathological cases which cannot occur in practice) are normally distributed. The discussion in this paper is not concerned with the "large sample case."

but simply their ranks, or perhaps only their signs. This feature of the methods arouses some criticism. It is intuitively obvious that a statistical procedure that replaces each of the two sets of numbers below by the *same* set of plus and minus signs:

$-8, -3, -2.1, -1.6, .3$	$- - - - +$
$-14, -14, -9.4, -.2, 5.7$	$- - - - +$

is not using all the information the data provide—is "throwing away" information. This is a less telling indictment than it seems to be. The relevant question is not "How much information does a certain statistical procedure throw away?", but rather "Of the methods available—classical, or non-parametric—which best uses the information in the sample?" Since the answer to the question will depend on the sort of population from which the sample comes, no general answer can be given. In the literature of mathematical statistics (**3, 23, 26**) there are examples of distributions where a non-parametric test which "throws away information" is clearly superior to a *t*-test, for instance. How the comparison would work in any given case is a matter of conjecture. The following advantages and disadvantages of the non-parametric methods should be considered:

Advantages of non-parametric methods:

1. Whatever may be the form of the distribution from which the sample has been drawn, a non-parametric test of a specified significance level actually *has* that significance level (provided that the sample has been drawn at random; in certain cases as will be noted, it is also necessary to assume that the distribution is continuous).

2. If samples are very small, e.g., six, there is in effect no alternative to a non-parametric test (unless the parent distribution really is known).

3. If the sample consists of observations from several *different* populations there may be a suitable non-parametric treatment.

4. The methods are usually easier to apply than the classical techniques.

5. If the data are inherently of the nature of ranks, not measurements, they can be treated directly by non-parametric methods without precariously assuming some special form for the underlying distribution.

6. In certain cases data can only be taken as "better" or "worse," that is, an observation can only be characterized as a plus or minus. Obviously, the classical tests are not directly applicable to such data.

Disadvantages of non-parametric methods:

1. If non-parametric tests rather than normal-theory tests are applied to normal data then they are wasteful of data. The degree of wastefulness is measured by the "efficiency" of the non-parametric test. If, for example, a test has 80 per cent efficiency this means that *where the data are from a normal distribution,* the appropriate classical test would be just as effective with a sample of 20 per cent smaller size. The efficiency thus expresses the relative

merits of the non-parametric test and the classical test under the conditions *where the normal test is correct*, but does not tell us how the tests will compare on non-normal data.

2. The non-parametric tests and tables of significance values are widely scattered in the periodical literature.

3. For large samples some of the non-parametric methods require a great amount of labor, unless approximations are employed.

Tests Based on Plus or Minus

The Sign Test. One of the best known and most widely applicable of the techniques to be discussed in this paper is the statistical sign test. A complete treatment will be found in (2). In many cases where an experimenter wishes to establish that two treatments are different (or that a particular one of the two is better) he is able to employ matched pairs, one member of each pair being assigned (at random) to treatment A, the other to treatment B. The classical technique is to apply a t-test to the differences; the underlying assumptions are that the differences are normally distributed with the same variance. The assumptions underlying the sign test are simply: (*a*) that the variable under consideration has a continuous distribution and (*b*) that both members of any pair are treated similarly—except for the experimental variable. There is an assumption neither of normality nor of similar treatment of the various pairs. Thus the different pairs may be of different socio-economic status, age, IQ, etc., so long as within each pair such relevant extranea are comparable. The hypothesis tested is that "The median difference is zero."[3] The test is performed by considering the differences $X_{Ai}-X_{Bi}$ and noting whether the sign is plus or minus. If the null hypothesis is true we expect about an equal number of plus and minus signs. The hypothesis is rejected if there are too few of one sign. The probability level of any result can be evaluated by the binomial expansion with $p=\frac{1}{2}$ and $N=$the number of pairs. Tables of significance values for various sample sizes are available (1). A table of sample sizes necessary to detect with probability .95 a departure from the null hypothesis of various degrees (e.g., that $P(X_A>X_B)=.3$) at significance levels .01, .05, .10, .25 is given by Dixon and Mood (2). For $N\geqq 30$ the normal approximation to the binomial will suffice.

If it is desired to test not merely that treatments A and B differ,

[3] The hypothesis is also properly expressed:

$$P(X_A > X_B) = P(X_B > X_A) = \tfrac{1}{2}$$

This is read in words: the probability that X_A will exceed X_B is equal to the probability that X_B will exceed X_A (and thus equal to $\frac{1}{2}$). X_A and X_B are two members of a pair.

but that treatment A is actually better than treatment B, a significant result can arise only if the number of minus signs is too small.

An extension of the sign test will permit one to determine whether A is better than B by, say, 5 points. Formally the null hypothesis is: $P(X_A > X_B + 5) \leqq \frac{1}{2}$.[4] In this case one considers the differences $X_{Ai} - (X_{Bi} + 5)$ and rejects the null hypothesis for too few minus signs.

Another extension enables one to determine whether A is better than B by some specified percentage—say 10 per cent. The null hypothesis is: $P(X_A > 1.10X_B) \leqq \frac{1}{2}$. If a significantly small number of the differences $X_{Ai} - (X_{Bi})$ (1.10) are negative the null hypothesis is rejected. Both these extensions are applicable only where the numbers are additive and the second is legitimate only if there is a zero point on the scale.

The efficiency of the sign test (in the sense defined) declines from around 95 per cent for $N = 6$ (25) to 62 per cent for very large samples. Where data are easily gotten, the extraordinary simplicity of computation sometimes justifies taking a larger sample and using the sign test, even though the classical methods would be justified and more efficient. In certain cases there is no substitute for the sign test, as where a pair of protocols can be assessed as to which exhibits more "cooperation" but there is little hope of a numerical evaluation.

The Median Test. In some cases where two treatments (or groups) are to be compared as to whether they are drawn from populations having the same median (or to determine whether a particular one of the two populations has a smaller median), it is not possible to work with matched pairs. The hypothesis can be tested by the median test (16, p. 394). The samples need not be of equal size. Suppose there are n X's and m Y's. Compute the median for the combined sample of $n+m$ observations. If the samples do come from populations with the same median then we should expect about half of the X's to be above the common median and about half below, similarly for the Y's. If the relative proportions are too discrepant, we reject the hypothesis of equality.

To perform the test, record a plus for any observation above the common median, a minus for any observation below the median. Then construct a 2×2 contingency table. For instance, suppose that an experiment yielded the following data:

Control Group: (X) 10, 15, 13, 12, 12, 14, 11, 9
Experimental Group: (Y) 7, 7, 8, 6, 13, 9

[4] In words this is read: the probability is at most $\frac{1}{2}$ that X_A will exceed X_B by 5. The null hypothesis can also be expressed: the median difference $X_A - X_B$ is equal to at most 5 in the population.

All observations greater than or equal to 11 are +'s; all 10 or less are −'s.

	+	−
Experimental Group	1	5
Control Group	6	2

The significance of the data is evaluated in the same manner as if this were a 2×2 test of independence. For such small frequencies as these Fisher's exact method must be used (5, Sec. 21.02); for large enough frequencies χ^2 with one degree of freedom is the test statistic, Yates's correction being used unless the number of cases is large. If the hypothesis is being tested against an alternative on one side only, i.e., the question asked of the data is not "are the two medians equal," but "is $\mathrm{Md}(X) \geqq \mathrm{Md}(Y)$," the ordinary techniques associated with χ^2 and one-sided test apply.

The assumptions underlying this test are that the X's and Y's are random samples from their respective populations, and that the population distributions are of the same form, differing only by a translation up or down the scale. Although the test is derived using the second assumption, Mood states that the test "is sensitive primarily to differences in location and very little to differences in shape."

Tests Based on Rank Order

There is a group of important methods which deal with the data in terms of their ranks. Four of the most important will be discussed here: a rank test for matched pairs (27, 28); the "T" test of Wilcoxon for two unmatched samples (27, 28), together with its extension by Mann and Whitney (13); the analysis of variance by ranks (6); the run test.

Wilcoxon's Matched Pairs Signed Ranks Test. Where the experimenter has paired scores X_{Ai} under treatment A and X_{Bi} under treatment B, he can rank the differences *in order of absolute size;* he may be unable to give numerical scores to the observations in each pair and still be able to rank the differences in order of absolute size. The ranking is done by giving rank 1 to the numerically *least* difference, rank 2 to the next least, etc. If methods A and B are equivalent, that is, if there is no difference and the null hypothesis is true, he should expect some of the larger, and some of the smaller, absolute deviations to arise with A being superior, some with B superior. That is, the sum of the ranks where A is favored should be about equal to the sum of the ranks where

B is favored. If the sum of the ranks for the negative differences is too small, or if the sum of the ranks for the positive differences is too small, the null hypothesis is to be rejected. Tables of significance values for the smaller sum of ranks will be found in (27) for n (the number of pairs) equal to 7 through 16. Tables for n from 7 to 25 are available in (29). For $n \geqq 25$ the sum of ranks T may be taken as normally distributed with mean $= \overline{T} = n(n+1)/4$ and standard deviation $\sqrt{(2n+1)\overline{T}/6}$. For example: suppose that seven pairs of rats are divided into a control and an experimental group. Suppose that the data are their times to run a certain maze and are as shown in Table 1.

TABLE 1

ILLUSTRATIVE DATA FOR TEST OF SIGNIFICANCE USING WILCOXON'S MATCHED PAIRS SIGNED RANKS TEST

Pair	*Exp.*	*Control*	*Diff. Exp-Control*	*Rank Diff.*	*Ranks with Less Frequent Sign*
(a)	65	51	14	6	
(b)	60	44	16	7	
(c)	71	64	7	4	
(d)	52	55	−3	1	1
(e)	62	49	13	5	
(f)	43	38	5	2	
(g)	58	52	6	3	
Total of ranks with less frequent sign					1

First, it is worth noting that these data are amenable to treatment by the sign test. Six of the differences have the same sign. The probability of six or more signs alike, if in fact the median difference is zero, is equal to $16/128 = 1/8$. Therefore, these data would not be regarded as cause for rejection using the sign test. But a closer examination of the data shows not only that there was only one negative difference but that it was the smallest difference in the set. These data argue more strongly against the null hypothesis than would the same set of differences with, say, pair (*e*) being the sole negative difference (or indeed any other one difference) though any of these possible samples would be treated identically by the sign test. It turns out that application of the rank test under consideration will adjudge these data as significant; essentially the different answer arises from exactly the considerations just sketched—the *size* of the sole negative difference is taken into account.

Wilcoxon's tables tell us that for $n=7$ a rank total of 2 or less for one of the groups is significant at level .05, and the null hypothesis of equality of treatments is rejected. The tables referred to are for two-sided tests. If one desires to test a one-sided hypothesis he may use the .05 level to determine a test of significance level .025, provided that the observed values lie in the direction of rejecting the one-sided hypothesis. Similar remarks apply to other significance levels.

A confidence interval for the difference in the treatment effects can be obtained as follows. Suppose that to the time of every rat in the control group $4\frac{1}{2}$ seconds were added, then all the differences would have the same sign as at present, the ranks would be the same, and the treatments would still be adjudged as significantly different. However, if $5\frac{1}{2}$ seconds were added to each control group score the groups would not differ significantly. The boundary for this argument occurs at 5. Similarly, if $14\frac{1}{2}$ be added to all the control group times the differences become all negative except for pair (*b*) which is then $+1\frac{1}{2}$ having a rank of 2.5 (it is tied with (*e*) for second and third place); this gives a "smaller rank total" of 2.5 which is not significant. But if 14 2/3 were added to each control group score then (*b*) would be the lone positive difference with rank 2; this would be significant. Since alterations in the differences greater than 5 and less or equal to 14.5 do not yield a significant difference, but values outside this range do, we can take 5 to $14\frac{1}{2}$ as a 95 per cent confidence interval for the increase in running time associated with the experimental treatment.

Mann-Whitney "U" Test. Where the observations are not made on matched pairs, but two unmatched groups are to be compared, the Mann-Whitney "*U*" test (or in the case of equal sized groups, its equivalent, the Wilcoxon "*T*" test) for two samples can be applied.

The null hypothesis which is tested is that the two groups of observations—say n X's and m Y's—have been drawn from a common population (that is, "there is no difference"). The test is designed to detect (roughly stated) whether one population has a larger mean than the other. Precisely stated, it is designed to guard against the alternative hypotheses that for every a, $P(X>a) \geqq P(Y>a)$ or $P(X>a) \leqq P(Y>a)$. A special case (unnecessarily restrictive) is where X and Y are assumed to have the same distribution except for a translation along the scale, so that the X's are all smaller—or all larger—than the "corresponding" Y's; here the null hypothesis says that there is no translation at all, and the test has the property that if in fact there is a positive or negative translation, then with a sufficiently large sample the test will reject the null hypothesis with any desired degree of probability.

To apply the test one arranges the $m+n$ observations in increasing order of size (algebraic sign not being ignored) and substitutes their ranks (1 for the smallest, $m+n$ for the largest). If the two samples were of equal size, so that $m=n$, the sum of the ranks for the X's should about equal the sum of the ranks for the Y's under the null hypothesis. If $m \neq n$ then the sums would be roughly proportional to the sizes m and n. The test consists in determining whether the observed discrepancy is too large to have arisen reasonably by chance, with the null hypothesis being true.

Tables of significance values for all possible pairs of sample sizes with $m \leqq 8$, $n \leqq 8$ are given in (13). For m and n both greater than 8 the test statistic is nearly normally distributed and the test of significance is made by employing this fact. If $m, n \geqq 8$, then U is normally distributed with mean $mn/2$ and standard deviation $\sqrt{mn(n+m+1)/12}$; one has merely to rank the $m+n$ observations from least to greatest, find T, the sum of the Y ranks, and from this calculate U, and see whether it is too many standard deviations removed from its expected value, $mn/2$.

TABLE 2

ILLUSTRATION FOR THE MANN-WHITNEY "U" TEST

Variable	*Observation*	*Rank*
X	10.2	1
X	12.8	2
Y	13.4	3
X	13.5	4
X	16.0	5
Y	17.1	6
Y	17.3	7
X	18.0	8
X	18.2	9
X	19.0	10
Y	19.4	11
X	19.5	12
Y	21.3	13
Y	24.0	14

$\sum X$ ranks $= 51$
$T = \sum Y$ ranks $= 54$.

As an example, suppose that there were 8 X's and 6 Y's, so that $m=6$, $n=8$ and that the data arranged in order of size were as shown in Table 2. The tables of significance are given in terms of U where

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

Here $U = 6 \times 8 + (6 \times 7/2) - 54 = 15$.

The table for $n = 8$ tells us that a U as small as 15 has a probability level of .141; so that the null hypothesis is accepted.

In using these tables the reader will find that the probability of small values of U is given. To find the probability $U \geqq k$ where k is a number larger than those given in the tables he uses the identity:

$$P(U \geqq k \mid n X\text{'s}, m Y\text{'s}) = P(U \leqq mn - k \mid n X\text{'s}, m Y\text{'s}).$$

It is further to be noted that m and n are entirely symmetrical, so that $P\{U = k \mid n\ X\text{'s}, m\ Y\text{'s}\} = P\{U = k \mid m\ X\text{'s}, n\ Y\text{'s}\}$.

As an example, suppose that an experimenter has 5 X's and 8 Y's and that the sum of the Y ranks, T, is 39. Then

$$U = 6 \times 8 + \frac{8(8+1)}{2} - 39 = 45.$$

This is a large value of U and is not tabled; to decide whether or not it is significantly large we note that

$$P(U \geqq 45) = P(U \leqq 6 \times 8 - 45) = P(U \leqq 3).$$

The tables tell us that this probability is .002.

Analysis of Variance with Ranked Data. The assumptions underlying the analysis of variance are: the observations are independent; they are drawn from normal populations all of which have the same variance; the means in these normal populations are linear combinations of "effects" due to row and/or columns, etc., that is, effects are additive.

Correlation among the observations would be perhaps the most dangerous assumption failure; but careful design should usually eliminate this. In some cases both normality of distribution and homogeneity of variance can be approximated either in the data, or by some transformation. In other cases this cannot be done. The analysis of variance by ranks is a very easy procedure and does not depend on such assumptions. It has the further advantage of enabling data which are inherently only ranks to be examined for significance.

Let there be n replications of an experiment where each subject undergoes a different one of p treatments. In each replication there are

a different p subjects. Data from such an experiment might be as follows:

TABLE 3

ILLUSTRATION FOR ANALYSIS OF VARIANCE WITH RANKED DATA

	Treatment				
	A	*B*	*C*	*D*	*E*
Group 1	11(2)	14(4)	13(3)	9(1)	20(5)
Group 2	12(3)	11(2)	13(4)	10(1)	18(5)
Group 3	16(3)	17(4)	14(2)	13(1)	19(5)
Group 4	9(1)	11(3)	14(4)	10(2)	16(5)
Rank totals	9	13	13	5	20

The numbers appearing in parentheses are the ranks from least to greatest *within each row* (replication). If the treatments A, B, C, D, E ($p=5$) are not different, then the rank totals would be expected to turn out about equal. In the present example there seems to be a marked disparity. To evaluate its significance we compute the statistic χ_r^2, done below, which has approximately the χ^2 distribution with $p-1$ degrees of freedom.

$$\chi_r^2 = \frac{12}{np(p+1)} \times \text{Sum (rank totals)}^2 - 3n(p+1)$$

Here $n=4$, $p=5$ and the statistic becomes:

$$\chi_r^2 = \frac{12}{120}\cdot(844) - 12(6)$$

$$= 12.4$$

For 4 degrees of freedom this is significant at level .02 but not .01.

If the groups 1, 2, 3, 4 in the example themselves represented four treatments, or age levels, etc., then a test of the equality of those four treatments could also be made by interchanging rows and columns. For that test χ_r^2 would have 3 degrees of freedom since then $p=4$, $n=5$.

A full treatment of the mathematical basis for the test is given by Friedman (6). Kendall and Smith (12) give exact probabilities for small m and n, and a detailed consideration of the closeness of approximation and recommendations for evaluation of significance levels are given in

Friedman's article (7). Wilcoxon (29) gives several instructive illustrations showing, among other things, how interactions can be tested.

Wald-Wolfowitz Run Test. The final test employing ranked data which will be considered is the Wald-Wolfowitz run test. This is a test of the hypothesis that two samples (not necessarily of equal size) have been drawn from a common population. It has the property that if the X's and Y's are not from a common population then, no matter in what way the populations differ (dispersion, median, skewness, etc.) the test will—for sufficiently large samples—reject the null hypothesis with probability as near to 1 as is desired. The application of the test is extremely simple.

Just as for the U test, arrange the combined sample of m Y's and n X's in increasing order. Then a run is defined as a sequence of letters of the same kind which cannot be extended by incorporating an adjacent observation. Thus there are 9 runs below:

$$\underline{X_1\ X_2}\ Y_1\ \underline{X_3}\ Y_2\ Y_3\ Y_4\ Y_5\ \underline{X_4\ X_5}\ Y_6\ \underline{X_6\ X_7\ X_8}\ Y_7\ \underline{X_9\ X_{10}}$$

The X runs are underlined; the Y runs stand between them.

Now if the two samples are from a common population then the X's and Y's will generally be well mixed and the number of runs will be large. But if the X population has a much higher median, then there is to be expected a long run of Y's at one end, a long run of X's at the other, and consequently a reduced total number of runs. If the X's come from a population with much greater dispersion then there should be a long run of X's at each end, and a reduced total number of runs. Similar arguments apply to opposite skewness, etc. Generally, then, rejection of the null hypothesis will be indicated if the runs are too few in number. An important application of the run test is to test randomness of grouping; in some such cases *either* too many or too few runs might be basis for rejection. A nice example is given by Swed and Eisenhart (24) where the question at issue is, are seats at a lunch counter a half hour before the rush hour occupied at random? Very many runs of occupied and empty seats would clearly be an a priori cause for rejection. So would too few runs if the possibility of friends coming together was to be considered. In the example to which the U test was earlier applied, only too few runs would be reasonable cause for rejection if the X's and Y's represented, say, examination scores for two different statistics classes.

The run test can also be applied to a series of events ordered in time. Let there be n observations arranged in order of the time at which they were taken. Let those greater than the median be denoted by X, those

less than the median by Y. If one suspects a time trend—like gradual increase—or a "bunching" in time due to change in attitude, etc., he would reject for too few runs.

Tables of significance for the run test are given by Swed and Eisenhart (24), for m, $n \leqq 20$. For larger samples the number of runs d can be taken as being normally distributed with mean $= (2mn/m+n)+1$ and

$$\text{standard deviation} = \sqrt{\frac{2mn(2mn - m - n)}{(m + n)^2(m + n - 1)}}.$$

Mood (16) states that for practical purposes this approximation will suffice for m, $n \geqq 10$. To apply this large sample theory one merely decides before taking the sample whether rejection is indicated by too many, or too few, (or either) runs and then sees whether d is too many standard deviations removed from its expected value in the rejection direction.

Mathematical investigations of this test indicate that because it guards against *all* kinds of difference between the distribution functions of X and Y it is not very powerful against any particular class of alternatives. Thus, if one were interested in detecting whether one population had a greater median than another he would do better to employ a test such as the U test. A related point is that when one rejects the null hypothesis on the basis of the run test, he can assert that the two populations differ—but he has little if any clue as to *how* they differ. Often the purpose is to establish that there is a difference in *means*, or *dispersion*, and the run test gives an answer which is not easy to interpret.

The only assumption involved in the run test is that the common population is continuous. This assumption is involved in all the tests depending on rank presented here. Generally, if there is a small number of ties the average rank for each set of tied observations may be given to each and the test carried through.

Randomization Tests

There is a variety of non-parametric tests which employ the numerical values of the data themselves. Among the most important of these are techniques based on the method of randomization. This kind of test was proposed by Fisher (4, Sec. 21), and has received extended treatment and development by Pitman (21, 22).

Matched Pairs. All the randomization tests are based on parallel logic. The simplest with which to exhibit the rationale is the matched

pair case. Suppose, for example, that we have two observations (one under condition A, the other under condition B) on each of seven individuals. The null hypothesis is that conditions A and B are no different; the data are shown in Table 4. The average difference is

TABLE 4

ILLUSTRATION FOR RANDOMIZATION TEST USING MATCHED PAIRS

i	X_{Ai}	X_{Bi}	$d_i = X_{Ai} - X_{Bi}$
1	14.9	15.5	−.6
2	17.3	16.5	.8
3	14.9	13.2	1.7
4	18.1	16.0	2.1
5	12.0	12.1	−.1
6	19.4	18.1	1.3
7	15.6	11.4	4.2
			9.4 = S

1.34, but is it significantly different from zero at, say, the 5 per cent level? To answer this with the t-test we would assume that the differences were normally distributed with a common unknown variance. We can get an exact test assuming only that the d_i are random samples from a common population. If the null hypothesis is true, then conditions A and B are experimentally indistinguishable, and for any individual the distinction between his X_A and X_B is merely a convention of labelling; in particular, the difference $X_{A3} - X_{B3} = 1.7$, say, is just exactly as likely as that $X_{B3} - X_{A3} = 1.7$. This means that associated with this sample are many other possible ones, all of which (under the null hypothesis) were exactly as likely to occur as this. For instance, the sample might just as well have turned out: +.6, −.8, −1.7, −2.1, +.1, −1.3, +4.2 or +.6, +.8, +1.7, +2.1, +.1, −1.3, −4.2, etc. In all, there are $2^7 = 128$ such outcomes, all equally likely under the null hypothesis that the treatments A and B are experimentally indistinguishable. With each of these is associated an $S = \Sigma d_i$. Some of these 128 S's are just about what one would expect if the null hypothesis were true, i.e., near zero. A few are well removed from zero—and much like what we expect under an alternative hypothesis such as the population mean of A exceeding that of B—or vice versa; we write these $\mu_A > \mu_B$ and $\mu_B > \mu_A$ in the sequel. To get an exact test of, say, level .05, we select of the samples which we can thus generate, that 5 per cent of them most likely under the alternatives we wish to guard against,

and constitute these chosen possible samples as our rejection region. In the present case, $.05(128) = 6.4$, so we choose six possibilities. The probability of getting one of these six samples under the null hypothesis is $6/128 = .047$. Then if the sample we actually drew is one of these listed for the rejection region we reject the hypothesis of equality of A and B. In our numerical example, if the investigator's "experimental hypothesis" had been: condition B leads to larger scores on the average than does condition A, he would test the null hypothesis of equality of A and B but would reject it only if the d_i were predominantly negative. If they were predominantly positive *or* well balanced he would have to regard the data as failing to support his experimental hypothesis. His rejection region would be six samples giving the greatest negative S. If he actually desired only to determine whether the two conditions yield different average scores then he must regard either a large positive S or a large negative S as cause for rejection, and his rejection region would consist of the three samples yielding greatest $+S$ and the negatives of these samples, which will yield the greatest $-S$.

Let us find the two-sided region just described. If all the d_i were positive then S would be 10.8, its maximum value. The next largest possible value for S (10.6) would be where all but $d_5 = .1$ were positive. Such considerations lead to the following list of the first 5 positive samples in order of size of S:

							S
.6	.8	1.7	2.1	.1	1.3	4.2	10.8
.6	.8	1.7	2.1	−.1	1.3	4.2	10.6
−.6	.8	1.7	2.1	.1	1.3	4.2	9.6
*−.6	.8	1.7	2.1	−.1	1.3	4.2	9.4
.6	−.8	1.7	2.1	.1	1.3	4.2	9.2

Thus the sample we obtained, which has been starred, lies in the acceptance region and the null hypothesis stands. If, however, only the alternative $\mu_A > \mu_B$ was being tested against, then the top six positive values would be the 5 per cent rejection region, and our sample, being 4th, would lie in it.

For large n, say 20, the number of possible samples which we can generate by altering signs on the given numbers is large ($2^{20} > 1{,}000{,}000$) and even listing a 1 per cent rejection region is a massive undertaking. There are two principal alternatives. Wilcoxon's T test, where ranks are substituted for numbers, may be used (in fact, the T test may be regarded as a randomization test on the ranks—and this clue should enable the reader to find the one-sided significance points for the T test where n is small). The second alternative is that where $n \geqq 12$

(roughly), and where the d_i are of roughly the same size (as a rule it might be safe to require $(d_k^2/\Sigma d_i^2) \leqq 5/2n$, where d_k is the largest difference in the set) a normal approximation can be used.

Each d_i, under the null hypothesis, is a chance variable taking the values $\pm d_i$ each with probability $\frac{1}{2}$. The d_i are independent. One form of the central limit theorem ensures that under the conditions given, the exact distibution of $z = S/\sqrt{\sum d_i^2}$ in the "randomization distribution" will be very closely approximated by the unit normal distribution. This test is obviously easy to apply. On data which are in fact normally distributed it is 100 per cent efficient for large samples. Examples of non-normal populations can be given where despite this efficiency it is an inferior test as compared with the rank T test (which has large sample efficiency of 95 per cent).

Two Sample Test. A randomization significance test for two samples has the same underlying logic. Let there be n X's and m Y's. If there is "no difference" then the fact that in the pooled ordered sample *a particular* n observations are labelled X is, so to speak, one of many equally likely accidents. All together there are

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

equally likely ways in which the relabelling might be done. For certain of these the "spread" or difference between ΣX and ΣY is extreme. The construction of the test consists in choosing a number k of these for a rejection region. If α is the significance level then k is chosen so that

$$k = \alpha \binom{m+n}{n},$$

as nearly as is possible. The choice of *which* k most extreme possible outcomes should constitute the rejection region depends, as always, on what alternatives are to be guarded against.

An example follows:

X	Y
11.6, 12.1, 12.2, 12.6, 13.1	9.5, 10.7, 11.8

We test $\mu_X = \mu_Y$ against the alternative $\mu_X > \mu_Y$.

The arithmetic is made more convenient if from all numbers we subtract 9.5, and then multiply by 10. We now have:

21 26 27 31 36	0 12 23

The average of these eight numbers is 22. If, then, the null hypothesis is true we should expect to find ΣY near (3) (22) = 66. In all there are

$$\binom{5+3}{3} = \frac{8.7.6}{1.2.3} = 56$$

possible equally likely samples. If we are working at level .05 we shall choose the 2 samples (out of the 56) most likely under the alternative hypothesis $\mu_X > \mu_Y$. These are, obviously

	23	26	27	31	36		0	12	21
and									
	21	26	27	31	36		0	12	23

The second of these is the sample we obtained, and the null hypothesis is rejected. For illustrative purposes the six most extreme (two-sided) samples are listed below:

X					Y		$\sum Y - 66$
23	26	27	31	36	0,	12, 21	$33-66=33$
0	12	21	23	26	27,	31, 36	$94-66=28$
21	26	27	31	36	0,	12, 23	$35-66=31$
0	12	21	23	27	26,	31, 36	$93-66=27$
21	23	27	31	36	0,	12, 26	$38-66=28$
0	12	21	26	27	23,	31, 36	$90-66=24$

If m and n are large, the carrying out of these computations becomes essentially impossible. But again there exists a convenient approximation to the distribution of the statistic in the randomization distribution of

$$\binom{m+n}{n}$$

possible sample values.

Provided that:

(1) $1/4 \leqq (m/n) \leqq 4$.

(2) $(\mu_4/\mu_2^2) - 3$ (The kurtosis computed for the pooled sample), not large; then the following statistic has approximately the t distribution with $m+n-2$ degrees of freedom:

$$\frac{\overline{Y} - \overline{X}}{\sqrt{\dfrac{\sum (Y - \overline{Y})^2 + \sum (X - \overline{X})^2}{m+n-2}\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$$

It is a curious result; this is the ordinary t statistic. This means that provided conditions 1 and 2 hold (and these can be checked from the

sample) the t statistic actually gives a test of the stated significance level *without* the usual assumptions being a part of the inference. It has *not* been assumed that X and Y are normally distributed with a common variance.

If the distributions of X and Y both have finite variance, but different means, the probability that the test will reject the null hypothesis tends to one as both m and n become large. If the distributions are different but have the same mean this is not so.

An alternative to the use of the t statistic to approximate the randomization distribution is to employ the Mann-Whitney U test. There are circumstances under which the U test (though it "throws away" data by reducing the observations to ranks) is the better test. The U test may be regarded as test of the randomization type applied to the *ranks* of the observations.

Confidence Intervals. In both these cases (paired or unpaired observations) confidence intervals can be obtained by adding equal increments to one set of values until a significant positive difference is first reached, and then altering them still further until a significant negative difference is first reached. These two extreme alterations constitute the end points of a confidence interval for the true difference. If the approximations (normal, and t) are to be used, then the conditions for their validity must hold at these extreme points; otherwise the exact procedure has to be used.

Correlation and Tests of Independence. The problem of correlation can also be attacked by the randomization method. That is, one can test the hypothesis of zero correlation with samples of small (or large) size without making assumptions about the form of the joint distribution of X and Y. For a treatment of the problem, see (22).

Tests of Independence

When one has a pair of observations (X_i, Y_i) for each member of his sample and desires to test the independence of X and Y there are numerous techniques available. The rank-order correlation coefficient, or τ, Kendall's rank-order statistic (11), may be used. The product-moment coefficient can be tested non-parametrically as mentioned in the preceding paragraph.

In addition there is an extraordinarily easily applied method, Olmstead and Tukey's corner test of association (20). Its efficiency and other properties await a full mathematical investigation, but informed opinion holds that it is likely to be a very good test.

To apply the test one first plots the observations in a scatter diagram. Then, following simple rules given by Olmstead and Tukey (20) and also by Mood (16), the statistician measures the degree to which the data are concentrated in the "corners" of the scatter diagram. (The instructions referred to essentially define the "corners.") If a substantial number of observations are concentrated in diagonally opposite corners (which would be expected in the presence of strong association between the variables), then the null hypothesis of independence is rejected. Although the use of this technique is simple, the explanation of how to construct the corners is rather lengthy and will be omitted here. The test is entirely distribution-free. Because of its ease of application it should find frequent use as a preliminary test to determine whether a product-moment correlation coefficient is worth computing in cases where the latter is fully justified.

There also exist non-parametric methods for linear regression, including tests of significance. They will not be taken up here, but a full treatment of both their mathematical theory and method of application is given by Mood (16, ch. 16). In this source there is also a technique for analysing one and two factor experiments; an alternative to the analysis of variance by ranks. All these methods depend upon the way in which the medians of various subclasses behave. They are all completely distribution free. As an attack on the analysis of variance problem they are more flexible than analysis of variance by ranks, but are less efficient, and probably not to be preferred for problems of an uncomplicated design.

Percentiles

If one has a sample of n observations and wishes to estimate the percentiles of the parent distibution he will, of course, employ the percentiles in the sample. Confidence intervals (confidence coefficient $1-\alpha$) may be obtained as follows. If the sample is arranged in order of increasing size:

$$X_1, X_2, X_3, \cdots, X_n$$

then X_1 is the smallest observation, X_2 the next smallest, etc. Let ξ_p denote the 100 p percentile. Then

$$P(X_r < \xi_p < X_s) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$$

Using tables of the binomial distribution (**19**), one then chooses r and s so that the probability (the value of the sum) is at least $1-\alpha$.

If there are ties in the data then

$$P(X_r < \xi_p < X_s) \geqq \sum_{i=r}^{s-1} \binom{n}{i} p^i - (1-p)^{n-i}$$

For example, a .90 confidence for the 40th percentile in the population from which this sample comes:

17	21	23	24	24	35	27	30
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8

is:

17 to 27,

that is:

X_1 to X_7.

Since:

$$\sum_{i=1}^{6} \binom{8}{i} (.4)^i (.6)^{8-i} = .90,$$

where we have $p = .4$, $n = 8$.

For large samples the binomial sum can be approximated by the normal distribution. The index i is approximately normal (for large n) with mean np and standard deviation $\sqrt{npq}$. So to obtain a 95 per cent confidence interval one would count $1.96\sqrt{npq}$ observations to the right and to the left of the $100p$ sample percentile to find the observations whose numerical values constitute a 95 per cent confidence interval.

Some Other Non-Parametric Methods

Certain important topics in the field of non-parametric methods have been either completely omitted, or merely mentioned in this paper. Among these are:

Rank Correlation Methods. A recent book by Kendall (**11**) provides the experimenter with a rather generous variety of techniques not elsewhere published. Among other matters of interest considered there are: tied ranks, coefficient of concordance (with significance test) to measure agreement among more than two judges, significance of the difference between two non-zero rank-order correlation coefficients. Work is being done in this field by Kendall and his associates and additional results will be published in the near future.

Kolmogorov-Smirnov Tests. These tests serve as alternatives (preferable for certain reasons) to χ^2 for two classes of problems:

1. To test the hypothesis that a random sample has been drawn from a population with a certain specified distribution.
2. To test that two random samples (of not necessarily equal size) have been drawn from the same population. The methods apply only where the chance variable is continuous. An excellent non-mathematical discussion, with tables and examples, is given by Massey (14). Some more recent results and tables for the two-sample problem are also given by Massey (15).

Tests for Randomness of a Sequence of Numerical Observations. The Wald and Wolfowitz run test discussed in this paper is one test of this sort, where two groups of observations are involved. Where there is only one sequence of observations, perhaps ordered in time, one may still wish to know whether they may be regarded as a random sequence. An informative non-mathematical discussion of this problem, with several tests, is found in Moore and Wallis (17).

Tolerance Intervals. One can ask the question: "Between what limits can I be nearly sure (say 95 per cent, or 99 per cent, etc.) that at least 90 per cent (or 80 per cent, or 98 per cent, etc.) of the population values lie?" These limits are called tolerance limits. The problem clearly differs from the confidence interval problem, which is concerned with location of the population mean, or a certain population percentile, etc.

A brief discussion will be found in Dixon and Massey (1). Some useful charts which eliminate computations are given by Murphy (18), where the relevant literature is also cited.

Literature on Non-Parametric Methods

The textbook literature presents few extended treatments of non-parametric methods. Of those known to the writer, one of the fullest, and surely the least mathematical, is Chapter 17 of Dixon and Massey's text (1). For the reader with facility in advanced calculus many important methods are explained and derived in Chapter 16 of Mood's text (16). At a mathematical level intermediate between these two is Chapter 8 of Johnson's text (9) and Chapter 9 of Hoel's text (8). Finally, the mathematically mature reader will find many of the techniques taken up in this paper (and some others) discussed in somewhat greater detail in Chapter 21, Volume II of Kendall's advanced book (10).

A paper by S. S. Wilks (30) affords a complete but terse review of the whole field up through about 1947. The treatment requires a good knowledge of mathematical statistics. A full bibliography is included.

The following bibliography is not intended to be complete. The reader who wishes to explore any one topic in detail will find little difficulty in uncovering the relevant literature with the aid of the references cited in the papers listed here.

BIBLIOGRAPHY

1. DIXON, W. J., & MASSEY, F. J., JR. *Introduction to statistical analysis.* New York: McGraw-Hill, 1951.
2. DIXON, W. J., & MOOD, A. M. The Statistical Sign Test. *J. Amer. statist. Ass.*, 1946, **41**, 557–566.
3. FESTINGER, L. The significance of difference between means without reference to the frequency distribution function. *Psychometrika*, 1946, **11**, 97–106.
4. FISHER, R. A. *Design of experiments.* London: Oliver and Boyd, 1936.
5. FISHER, R. A. *Statistical methods for research workers.* London: Oliver and Boyd, 1925.
6. FRIEDMAN, MILTON. Use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. statist. Ass.*, 1937, **32**, 675–701.
7. FRIEDMAN, MILTON. A comparison of alternative tests of significance for the problem of m rankings. *Ann. math. Statist.*, 1940, *11*, 86–92.
8. HOEL, P. G. *Introduction to mathematical statistics.* New York: Wiley, 1947.
9. JOHNSON, P. O. *Statistical methods in research.* New York: Prentice-Hall, 1949.
10. KENDALL, M. G. *The advanced theory of statistics*, Vol. II. London: C. Griffin and Co., 1948.
11. KENDALL, M. G. *Rank correlation methods.* London: C. Griffin and Co., 1948.
12. KENDALL, M. G., & SMITH, B. B. The problem of m rankings. *Ann. math. Statist.*, 1939, **10**, 275–287.
13. MANN, H. B., & WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. math. Statist.*, 1947, **18**, 50–60.
14. MASSEY, F. J., JR. The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. statist. Ass.*, 1951, **46**, 68–78.
15. MASSEY, F. J., JR. The distribution of the maximum deviation between two sample cumulative step functions. *Ann. math. Statist.*, 1951, **22**, 125–128.
16. MOOD, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.
17. MOORE, G. H., & WALLIS, W. A. Time series significance tests based on signs of differences. *J. Amer. statist. Ass.*, 1943, **38**, 153–164.
18. MURPHY, R. B. Non-parametric tolerance limits. *Ann. math. Statist.*, 1948, **19**, 581–589.
19. NATIONAL BUREAU OF STANDARDS. *Tables of the binomial probability distribution.* Washington, D. C.: U. S. Government Printing Office, 1949.
20. OLMSTEAD, P. S., & TUKEY, J. W. A corner test for association. *Ann. math. Statist.*, 1947, **18**, 495–513.
21. PITMAN, E. J. G. Significance tests which may be applied to samples from any population. *Suppl. J. Royal statist. Soc.*, 1937, **4**, 119.
22. PITMAN, E. J. G. Significance tests which may be applied to samples from any population, II. The correlation coefficient test. *Suppl. J. Roy. statist. Soc.*, 1937, **4**, 225.
23. PITMAN, E. J. G. Notes on non-parametric statistical inference. (Unpublished.)
24. SWED, F. S., & EISENHART, C. Tables for testing randomness of grouping

in a sequence of alternatives. *Ann. math. Statist.*, 1943, **14**, 66–87.

25. WALSH, J. E. On the power of the sign test for slippage of means. *Ann. math. Statist.*, 1946, **17**, 358–362.
26. WHITNEY, D. R., *A Comparison of the power of non-parametric tests and tests based on the normal distribution under nonnormal alternatives.* Unpublished Ph.D. dissertation at Ohio State University, 1948.
27. WILCOXON, FRANK. Individual comparisons by ranking methods. *Biometrics Bull.*, 1945, 1, 80–82.
28. WILCOXON, FRANK. Probability tables for individual comparison by ranking methods. *Biometrics*, 1947, **3**, 119–22.
29. WILCOXON, FRANK. *Some rapid approximate statistical procedures.* American Cyanamide Co., 1949.
30. WILKS, S. S. Order statistics. *Bull. Amer. math. Soc.*, 1948, **54**, 6–50.

Received July 19, 1951.