ARKIV FÖR MATEMATIK Band 3 nr 9

Communicated 10 March 1954 by Arne Westgren and Harald Cramér

Comparison of tests for non-parametric hypotheses

By Erik Ruist

With 1 figure in the text

1. Summary

The main object of this paper is to find a criterion for comparison of two tests for non-parametric hypotheses, taking advantage of the qualitative information that may exist. After a detailed analysis of the problem and some earlier suggestions for its solution (sections 2–4), a criterion is suggested in section 5. In order to apply it to a concrete case, a location problem is specified in section 6. The rank tests to be compared are analyzed in section 7, and the comparison by way of the criterion is carried out in section 8. It turns out that sign tests, sometimes slightly modified, are very often optimal according to the criterion used.

2. Statistical hypotheses testing

The general problem of testing statistical hypotheses can be described in the following way. We are interested in the distribution F of a stochastic variable X, which may generally be thought of as a vector $\{X_1, X_2, \ldots, X_n\}$ the elements of which may correspond to different dimensions of the variable and/or to several observations on the same variable. Let us make the following basic assumptions. In the sample space there is a σ-algebra of subsets $(\mathfrak{X})$. All probability measures are defined on $(\mathfrak{X})$. All functions introduced in the following are measurable $(\mathfrak{X})$, and all sets belong to $(\mathfrak{X})$.

We know that $F \in \Omega$, a class of probability distributions. After drawing a sample, i.e. observing a sample point x, we want to decide to which of a number of mutually exclusive sub-classes $\omega_1, \omega_2, \ldots, \omega_m$ of Ω we shall refer F. In the most common cases, $m = 2$. The problem is then often stated to be the testing of the null hypothesis H_1: $F \in \omega_1$, against the alternative hypothesis H_2: $F \in \omega_2$. If ω_i is not topologically equivalent to a finite-dimensional Euclidean space, H_i is said to be a non-parametric hypothesis.[1] In the following, we shall be mainly concerned with such hypotheses.

[1] The name is not a very good one, as ω_i is often characterized by a parameter. As an example, ω_i may contain all F the marginal distributions of which, with respect to all X_j, have positive medians. The term seems to have been first used by Wolfowitz (1942) for the truly non-parametric two-sample problem where ω_1 contains all F that are invariant with respect to all permutations of $X_1, \ldots, X_n$, while ω_2 contains all F invariant only to permutations within each of the sequences $X_1, \ldots, X_k$ and $X_{k+1}, \ldots, X_n$, but not between them.

Let us denote the decision to accept H_i by d_i ($i = 1, 2$). It is then required to construct a *test* φ, i.e. a function $\varphi(x)$ of the sample point x, such that $0 \leq \varphi(x) \leq 1$, indicating for every sample point the probability with which decision d_2 should be made. By the *size* of the test we shall mean

$$\alpha = \sup_{F \in \omega_1} P(d_2 \mid \varphi, F) = \sup_{F \in \omega_1} \int \varphi(x)\, dF(x)$$

where the integral is to be taken over the entire sample space. If $P(d_2 \mid \varphi, F) = \alpha$ for all $F \in \omega_1$ the test is said to be *similar* with respect to ω_1. Usually, with each test $\varphi(x)$ there is associated a real-valued function $\psi(x)$ such that

$$\varphi(x) = \begin{cases} 1 & \text{if } \psi(x) > c \\ p & \text{if } \psi(x) = c \\ 0 & \text{if } \psi(x) < c \end{cases} \tag{1}$$

where c and p are parameters, $0 \leq p \leq 1$. The choice of c and p determines the value of α. By changing c and p any value of α $(0 \leq \alpha \leq 1)$ can be obtained. To every value of α there is only one relevant combination of c and p (if $P\{\psi(x) = c \mid F \in \omega_1\} = 0$, the value of p is irrelevant). Thus $\psi(x)$ may be said to generate a *family of tests* $\Phi = \{\varphi(x)\}$ where to each value of α corresponds one, and only one, $\varphi(x)$. If necessary, we shall denote this member of the family by φ_α. According to this definition, the t test is a test family, generated by

$$\psi(x) = \frac{\bar{x} - m}{s}\sqrt{n-1}.$$

Another type of tests which is of interest in the present context is obtained if c and p are replaced by functions $c(x)$ and $p(x)$ of the sample point. This is the case with the invariance tests investigated by LEHMANN and STEIN (1949). Denote by $x^{(1)}, x^{(2)}, \ldots, x^{(M)}$ the sample points obtained from x by attaching in all possible ways plus and minus signs to the absolute values of the components of x, and let the order among the sample points be defined by

$$\psi(x^{(1)}) \geq \psi(x^{(2)}) \geq \cdots \geq \psi(x^{(M)}).$$

Then we define $c(x) = \psi(x^{(1+[M\alpha])})$ where $[M\alpha]$ denotes the largest integer less than or equal to $M\alpha$. The value of $p(x)$ is determined so as to give to the test the desired size α. A simple example is the following test of symmetry around 0: Let $\frac{1}{n}\psi(x) = \bar{x}$, the sample mean. Suppose that the sample values are 6, 3, -1, 4, 2, 3, so that $\psi(x) = 17$. If we choose $\alpha = 0.05$, we get $[M\alpha] = 3$. Further, $\psi(x^{(1)}) = 19$, $\psi(x^{(2)}) = 17$, $\psi(x^{(3)}) = 15$, $\psi(x^{(4)}) = 13$, so $c(x) = 13$, and since $\psi(x) > c(x)$ we find $\varphi(x) = 1$, i.e. we reject the hypothesis of symmetry.

For this type of tests also, any value of α could be obtained by the use of one $\psi(x)$. Thus, this function generates one family of tests when c is a parameter, independent of x, and another family when $c(x)$ takes values depending on x in the way described above. We may call them the fixed-limit test family and the permutation test family generated by $\psi(x)$. It is clear that any strictly increasing function of $\psi(x)$ will generate the same families as $\psi(x)$. HOEFFDING (1952) has shown that, subject to certain restrictions, the two families generated by a function $\psi(x)$ are asymptotically equivalent for increasing n. In this paper we shall be concerned with these types of tests exclusively, and by "a test" we shall mean a test belonging to a fixed-limit or a permutation test family.

A test family may be said to be *distribution-free* in a class ω of distributions, if the distribution of $\varphi(x)$ is the same for all $F \in \omega$. This means, in the cases considered here, that the probabilities $P\{\varphi(x)=1\}$, $P\{\varphi(x)=p\}$, and $P\{\varphi(x)=0\}$ are the same for all $F \in \omega$. Thus, the t test family is distribution-free in the class of normal distributions with mean m. When testing a non-parametric hypothesis H_1, it is often convenient to use a test from a family that is distribution-free in the corresponding ω_1. Accordingly, these tests are usually called non-parametric. As they are also used in other cases, and as they often test the value of some parameter, the name is misleading, and the term distribution-free will be used here.

There is a simple relation between similar tests and distribution-free test families as is seen from the following

Theorem 1. A test family is distribution-free in ω_1, if and only if all tests belonging to it are similar with respect to ω_1.

Proof: For the sake of clarity, denote c and p by c_α and p_α. Then, for any given α and F, let

$$\pi_1(\alpha, F) = P\{\varphi(x)=1 \mid \alpha, F\} = P\{\psi(x) > c_\alpha \mid F\}$$
$$\pi_2(\alpha, F) = P\{\varphi(x)=p_\alpha \mid \alpha, F\} = P\{\psi(x) = c_\alpha \mid F\}$$
$$\pi_3(\alpha, F) = P\{\varphi(x)=0 \mid \alpha, F\} = P\{\psi(x) < c_\alpha \mid F\}.$$

By definition, $\sum_{i=1}^{3} \pi_i(\alpha, F) = 1$ and $\pi_1(\alpha, F) + p_\alpha \pi_2(\alpha, F) = \alpha$ for all $F \in \omega_1$. Now suppose that for two distributions $F', F'' \in \omega_1$ the inequality $\pi_i(\alpha, F') \neq \pi_i(\alpha, F'')$ is true at least for some i and for some α, say α_1. Then define $\alpha_2 = \pi_1(\alpha_1, F')$ and

$$\alpha_3 = \pi_1(\alpha_1, F') + \pi_2(\alpha_1, F').$$

To reach these sizes for F', we must make

$$c_{\alpha_2} = c_{\alpha_3} = c_{\alpha_1}$$
$$p_{\alpha_2} = 0 \qquad p_{\alpha_3} = 1.$$

But, as the test is similar,

$$\pi_1(\alpha_2, F'') = \pi_1(\alpha_3, F'') = \pi_1(\alpha_1, F'') = \alpha_2$$
$$\pi_2(\alpha_3, F'') = \pi_2(\alpha_1, F'') = \alpha_3 - \alpha_2$$

and

$$\pi_1(\alpha_1, F') = \pi_1(\alpha_1, F'') = \alpha_2$$
$$\pi_2(\alpha_1, F') = \pi_2(\alpha_1, F'') = \alpha_3 - \alpha_2$$
$$\pi_3(\alpha_1, F') = \pi_3(\alpha_1, F'') = 1 - \alpha_3.$$

Thus, the probabilities must be equal for any value of α, and the test family is distribution-free, which completes the proof of the first part of the theorem. As the converse of this is trivial, the second part follows.

3. Optimality criteria

The choice among all possible tests in a given situation is guided by the general aim of keeping the probability of a wrong decision as small as possible. For a test φ_α, the probability of wrongly making decision d_2 is at most α, and the probability of wrongly making decision d_1 is $P(d_1 \mid \varphi_\alpha, F \in \omega_2)$. Its complement, $P(d_2 \mid \varphi_\alpha, F \in \omega_2) = 1 - P(d_1 \mid \varphi_\alpha, F \in \omega_2) = \int \varphi_\alpha(x)\, dF(x)$ is called the *power* of the test, and is a function of F. Several criteria have been suggested for discriminating between the tests. Some of these pick out one *test* as the best one, weighting in some way the seriousness of the different kinds of wrong decisions. Others point out the *test family* to be used, leaving the size of the test to be determined subjectively. These criteria are based on the power function only. Sometimes the choice of size is supposed to be made prior to the choice of a test family, but this is generally equivalent to using the reverse order, as in most cases all tests of a family have the same optimality properties.

Suppose first that the size of the test is chosen subjectively. Then in the simple case when ω_1 and ω_2 each contain only one distribution, say F_1 and F_2, it is clearly desirable to use a test family which for any given α maximizes $P(d_2 \mid \varphi_\alpha, F_2)$, i.e. *most powerful* tests. Even in more general cases there sometimes, though rarely, exists a *uniformly most powerful* test family, the tests φ' of which satisfy

$$P(d_2 \mid \varphi'_\alpha, F) = \max_{\varphi_\alpha} P(d_2 \mid \varphi_\alpha, F) \quad \text{for all } F \in \omega_2.$$

When no uniformly most powerful tests exist, the choice between several test families is not so obvious. If, however, two tests of size α satisfy

$$\begin{aligned} P(d_2 \mid \varphi'_\alpha, F) &\geq P(d_2 \mid \varphi_\alpha, F) \qquad \text{for all } F \in \omega_2 \\ P(d_2 \mid \varphi'_\alpha, F) &> P(d_2 \mid \varphi_\alpha, F) \qquad \text{for some } F \in \omega_2 \end{aligned}$$

φ'_α is said to be *uniformly more powerful* than φ_α (WALD, 1942), and φ'_α is clearly preferred to φ_α. A test for which no uniformly more powerful test exists may be called α-*admissible* (LEHMANN, 1947, called such a test admissible). A test that is more powerful than any other test of the same size for at least one $F \in \omega_2$ is obviously α-admissible. A test family all tests of which are α-admissible may be called admissible. Test families that are not admissible are usually undesirable.

For the choice between admissible test families some additional criterion, more or less subjectively chosen, is necessary. It may have the effect of disqualifying a class of test families having a property that is regarded as undesirable. An example of such a property is bias, which means that there exists a $F' \in \omega_2$ for which $P(d_2 \mid \varphi_\alpha, F') < \alpha$. If all biased families are excluded from the class of admissible test families, there may sometimes be only one left, the uniformly most powerful unbiased test family.

Another method of choosing between admissible families is to take special regard to the power against some specific $F \in \omega_2$. Thus, for the case where ω_1 contains only one element, and ω_2 is a parametric family of distributions, thus Ω being a finite-dimensional parameter space, NEYMAN and PEARSON (1936) suggested that among all unbiased tests of size α, the test with the steepest power function in the vicinity of ω_1 should be regarded as best. They knew that the criterion could be criticized, since usually a wrong decision d_1 is not so serious when F is close to ω_1 as when it is far from it. They hoped, however, that a good test according to this criterion would also behave well farther away from ω_1.

WALD (1942) suggested another criterion, saying that a test φ'_α is *most stringent*, if it minimizes

$$\max_{F\in\omega_2} [\sup_{\varphi_\alpha} P(d_2 \mid \varphi_\alpha, F) - P(d_2 \mid \varphi'_\alpha, F)].$$

Thus, the criterion takes account of the maximal deviation from the envelope power function.

From an intuitive point of view, an optimality criterion should preferably take account of the whole of the power function. LINDLEY (1953) tried to do this, saying essentially that a test is optimal, if it minimizes a weighted average probability of a wrong decision. Thus, he considered not only $P(d_1 \mid \varphi, F\in\omega_2)$ but also $P(d_2 \mid \varphi, F\in\omega_1)$ and accordingly he obtained a unique optimal test, not a test family. In the situation described above, if ω_2 is a one-parameter family $\{F(\theta)\}$, a test is optimal according to Lindley if it minimizes

$$v_1 P(d_2 \mid \varphi, F\in\omega_1) + \int_{\theta\in\omega_2} v_2(\theta) P(d_1 \mid \varphi, F(\theta))\, d\theta$$

where v_1 and $v_2(\theta)$ are fixed weights. The weights are derived by a combined evaluation of the belief in the plausibility of the distribution and the seriousness of committing the corresponding error. It seems very difficult to determine such weights in a given situation, and hence the applicability of the criterion is restricted.

WALD (1950) attacked the problem differently, but he also introduced numerical weights, first to all $F\in\omega_2$, indicating the seriousness of, or the loss suffered from, making decision d_1 when the true distribution is F. Considerations of symmetry led to the introduction of corresponding weights in ω_1, indicating for any $F\in\omega_1$ the seriousness of making the wrong decision d_2. Let us denote the weight function by $W(F, d_i)$. For completeness, let $W(F, d_i) = 0$ for $F\in\omega_i$ $(i = 1, 2)$. Now, for any φ and any F, the expected loss equals $r(\varphi, F) = r_1(\varphi, F) + r_2(\varphi, F)$ where

$$r_i(\varphi, F) = P(d_i \mid \varphi, F)\, W(F, d_i).$$

A test φ' is said to be *uniformly better* than φ if

$$\begin{aligned} r(\varphi', F) &\le r(\varphi, F) && \text{for all } F\in\Omega, \text{ and} \\ r(\varphi', F) &< r(\varphi, F) && \text{for some } F\in\Omega \end{aligned}$$

which is in this case equivalent to

$$\begin{aligned} P(d_i \mid \varphi', F) &\ge P(d_i \mid \varphi, F) && \text{for all } F\in\omega_i,\ i = 1, 2, \text{ and} \\ P(d_i \mid \varphi', F) &> P(d_i \mid \varphi, F) && \text{for some } F\in\omega_i. \end{aligned}$$

This concept is seen to be related to the concept "uniformly more powerful", but is more general, since it admits a comparison of two tests of different sizes, and stronger, since it takes account of ω_1 as well as of ω_2. A test for which no uniformly better test exists is called *admissible*. This is a stronger concept than α-admissibility. Admissibility implies α-admissibility, but the contrary is not necessarily true.

For the choice between admissible tests, WALD introduced the *minimax principle*, saying that the test φ' is optimal in the minimax sense if

$$\sup_{F \in \Omega} r(\varphi', F) = \min_{\varphi} \sup_{F \in \Omega} r(\varphi, F).$$

By this criterion also, we do not get a family of tests of different sizes, but usually only one test, owing to the fact that we have evaluated the relative seriousness of committing the two possible kinds of error. According to a theorem of SVERDRUP (1952), a minimax test is admissible, if it is unique, but not necessarily otherwise.

For any test φ, the distribution F' for which

$$r(\varphi, F') = \sup_{F \in \Omega} r(\varphi, F)$$

is, if it exists, called a least favorable distribution with respect to φ. WALD showed that, subject to rather mild restrictions, the minimax test φ' satisfies

$$r(\varphi', F') = \min_{\varphi} r(\varphi, F')$$

if F' is a least favorable distribution with respect to φ'. Thus, a minimax test is optimal for its least favorable distributions, meaning inter alia that if $F' \in \omega_2$, φ' is more powerful relative to F' than any other test of the same size.

Let, as above, ω_1 contain only one element and ω_2 be a parametric family, and $W(F, d_i) = k_i$ for $F \notin \omega_i$. Then, for all unbiased tests,

$$r_1(\varphi, F) < k_1(1-\alpha),$$

but if the power function is continuous, $r_1(\varphi, F) \to k_1(1-\alpha)$ as $F \to \omega_1$. Thus, the maximum risk in ω_2 is $k_1(1-\alpha)+\varepsilon$ where $\varepsilon \to 0$. In ω_1 the risk r_2 is always $= k_2\alpha$. Then, a test is minimax if it is optimal for $F \to \omega_1$, and of size α so that $k_1(1-\alpha) = k_2\alpha$, i.e. $\alpha = \dfrac{k_1}{k_1+k_2}$. Such a test has higher power in the close vicinity of ω_1, i.e. steeper power function in that vicinity, than any other test, and is thus optimal according to the Neyman-Pearson criterion. This is a slight modification of a theorem by SVERDRUP (1953, p. 85).

The minimax approach is applicable to any set Ω and is thus of great generality. It has, however, been subjected to some criticism. Thus, it is said not to be practical to let the worst possible situation guide one's behavior. Nor is it usually possible to define the absolute value of the loss for different F, i.e. the weight function. Specifically, the effect of wrongly making decision d_1 may be of a kind quite different from that of wrongly making decision d_2. Thus, the choice of a test within the "best" test family cannot be made on the basis of quantitative judgement. The practical applicability of the minimax principle to non-parametric problems is limited also by the fact that the problem of finding the minimax test for a given situation has so far been solved only for some special cases.

HOEFFDING (1951) has used the minimax criterion in a way slightly different from WALD's. Thus, he agrees with the second part of the criticism cited above, that it is often impossible to introduce a weight function in Ω, and to compare the two kinds of wrong decisions. Consequently, he restricts himself to defining a weight function $W(F, d_1)$ in ω_2 only. We might call such a weight function *incomplete*. HOEFFDING defines a test φ'_α of given size α as being of minimax risk if

$$\sup_{F\in\omega_2} r_1(\varphi'_\alpha, F) = \min_{\varphi_\alpha} \sup_{F\in\omega_2} r_1(\varphi_\alpha, F).$$

He further points out that although it may not be realistic to give numerical weights to different $F\in\omega_2$, it is usually possible to rank all pairs F_1, $F_2\in\omega_2$ according to whether the weight of F_1 should be greater than, equal to, or smaller than that of F_2. This is equivalent to partitioning ω_2 into disjoint sets $\Omega(a)$ where a is a real valued parameter such that

$$W(F_1, d_1) \le W(F_2, d_1) \text{ if } F_1\in\Omega(a_1) \text{ and } F_2\in\Omega(a_2) \text{ where } a_1 < a_2,$$

$$\text{and } W(F_1, d_1) = W(F_2, d_1) \text{ if } F_1, F_2\in\Omega(a).$$

We may call such a weight function *a-ordered* incomplete.

Now, HOEFFDING has shown that if a test φ'_α satisfies the condition

$$\inf_{F\in\Omega(a)} P(d_2 \mid \varphi'_\alpha, F) = \max_{\varphi_\alpha} \inf_{F\in\Omega(a)} P(d_2 \mid \varphi_\alpha, F) \text{ for all } a$$

it minimizes, among all tests of size α, the maximum risk $r_1(\varphi_\alpha, F)$ with respect to all a-ordered incomplete weight functions. Thus, HOEFFDING's criterion gives tests that are optimal in a great many situations, but it does not, like WALD's, choose a test, only a test family. HOEFFDING gives examples of such test families of *maximin power with respect to* $\{\Omega(a)\}$. As in the case of complete weight functions, we may speak of $F'\in\omega_2$ as a least favorable distribution in ω_2 relative to a test family Φ, if for any α

$$r_1(\varphi_\alpha, F') = \sup_{F\in\omega_2} r_1(\varphi_\alpha, F).$$

By restricting F to $\Omega(a)$ we get by the same definition a least favorable distribution in $\Omega(a)$. Since $W(F, d_1)$ is constant within $\Omega(a)$, the condition is in this case equivalent to

$$P(d_2 \mid \varphi_\alpha, F') = \inf_{F\in\Omega(a)} P(d_2 \mid \varphi_\alpha, F).$$

So far there is no general way of establishing the existence of such a test family of maximin power in a given situation, nor of finding it, if it exists. Thus, even after the introduction of HOEFFDING's criterion, it is not always possible to find an optimal test. This is especially true when Ω is a non-parametric class of distributions, i.e. when little or nothing is known of the functional form of F.

The main results in finding optimal tests with respect to non-parametric classes of distributions are due to LEHMANN and STEIN (1949). They found that if ω_1 contains all distributions invariant with respect to permutations of the variables $X_1, X_2, \ldots, X_n$, and ω_2 contains only one element, say F', the most powerful test family is a permutation family with $\psi(x) = f'(x)$, the density function of F'. By a theorem of HUNT and STEIN, they extended the results to some cases where ω_2 is a parameter space. The results are very important, as the invariance hypothesis tested is a very common one, including as special cases the two-sample problem and the problem of symmetry.

However, the computational work necessary for the tests obtained is prohibitive unless the sample is very small, say less than 10. Further, the tests' property of being uniformly most powerful is retained only as long as ω_1 is a very wide class of distributions.

Thus, in this case as well as in others where no optimal test has been found, it is necessary to use less powerful tests, and to decide which of two given tests is to be considered the better.

4. Comparing two tests or test families

Of course, the criteria referred to above are capable of picking out the better of two tests as well as of finding a best test. Thus, if one test is uniformly more powerful than another, it should generally be preferred to the second one. If none of the tests is uniformly more powerful than the other one, we have to apply some other criterion, which is sometimes difficult if Ω is not a parametric space. For several distribution-free tests, such comparisons have been made, based on different criteria, the choice between these generally depending more upon what is possible than on what is desirable. To understand the background of these comparisons, it is necessary to remember under what circumstances distribution-free tests are being used.

Besides being distribution-free, these tests have often the property of being very simple to compute, and that is the reason why they are applied in two distinctly different situations:

1. When the number of observations is large, and computational ease may be worth some loss of power. Even if the form of the distribution is unknown, most of the standard tests could have been applied in this case, as the $\psi(x)$ corresponding to them are often asymptotically normally distributed for almost any F.

2. When the number of observations is small, and the functional form of the distribution is unknown. The standard tests could not be applied, as the sample size is too small for the asymptotic normality to be relied upon. This is very often the case in statistical problems relating to the social sciences.

For the first case, it is of course of interest to see how much power is lost by not using the most powerful test. It is quite enough for this purpose to make the investigation under the assumption that F is normal in some sense, and that the number of observations is so large that asymptotic results may be used. The comparison is ordinarily made in terms of relative efficiency. The *efficiency* of test φ'_α relative to φ''_α is said to be e, if the power of φ'_α using n observations is in some sense equivalent to the power of φ''_α, using $e\,n$ observations. If φ''_α is a uniformly most powerful test, e may be called the absolute efficiency of φ'_α. The first to compute the efficiency of a distribution-free test seems to have been COCHRAN (1937), who used a sign test (see below, p. 147) for testing Student's hypothesis, and found its efficiency (relative to the uniformly most powerful unbiased t test) to tend to $2/\pi$ as n increased infinitely. By his definition the powers are equivalent if their slopes are equal in the near vicinity of ω_1. This measure of efficiency is of course related to the Neyman-Pearson definition of an optimal test, referred to above, and may be called *asymptotic local efficiency* (BLOMQVIST, 1950). The same measure was obtained for Wilcoxon's one-sample test (cf. p. 147) by PITMAN (1948), who found it to be $3/\pi$. Thus, according to Neyman-Pearson's criterion, Wilcoxon's test should be preferred to the sign test when ω_2 contains only normal distributions and the sample is large. PITMAN applied the same technique to Wilcoxon's two-sample test, whose asymptotic local efficiency relative to the t test was found to be $3/\pi = 0.96$ if ω_2 contained normal distributions, 1 for rectangular distributions, and 81/64 for a $\Gamma(3)$ distribution, giving a rough indication of the applicability of the test for big samples.

It was felt, however, that the local efficiency of the distribution-free tests decreases when the sample size increases, and thus it was of interest to compute the relative local efficiency of some tests also for small samples. VAN DER VAART (1950) computed the first two derivatives of the power function of the Wilcoxon two-sample test for ω_2 normal and $n \leq 6$, and compared them with the corresponding measures for the t test with the same number of observations. He did not work out the value of the relative efficiency as defined above, but approximate calculations based on his results indicate that the local efficiency of Wilcoxon's test relative to the t test is at least 0.95 for these sample sizes.

However, it seemed necessary to look not only for local properties of the tests, but for more general properties of the power functions. WALSH (1946, 1949) made extensive calculations for different tests when Ω is a class of normal distributions, and when the power for different $F \in \omega_2$ depends only on one parameter, say θ. He chose to introduce a new definition of equivalence of power functions and obtained what may perhaps be called *average efficiency*. According to WALSH's definition, two power functions are equivalent if the integral of their difference over the range of θ corresponding to ω_2 is equal to 0. This definition is obviously closely related to LINDLEY's optimality criterion, except that WALSH gives the same weight to all $F \in \omega_2$.

In the examples given above, the power efficiency in some sense was computed for very restricted ω_2, mainly normal ones. Where, moreover, only asymptotic expressions were derived, they point mainly to the usefulness of the tests in case 1, and may serve as rough guides only for case 2.

Without assuming normality or some other functional form of the distribution, very few results on the power of distribution-free tests exist. HOEFFDING (1952) has shown that under certain restrictions the tests belonging to the permutation family generated by a function $\psi(x)$ have asymptotically the same power against a certain ω_2 as the corresponding tests of the fixed limit family generated by $\psi(x)$.

Sometimes, it is possible to define a partition $\{\Omega(a)\}$ such that the power for certain tests is constant within every $\Omega(a)$. If so, the methods used for parametric problems can be directly applied. LEHMANN (1953) found such a partition and a corresponding class of tests for the two-sample problem. Thus, within this class the tests may be compared using, say, WALSH's method. It is, however, rarely possible to find such a partition. Thus, there does not seem to be any attempt so far to develop a general method for comparing two distribution-free tests when the reason for their use is the second one listed above. Curiously enough, little has been done even for the development of optimal tests except on a very general or a very specified level. This may be the reason for the common complaint that too much information is lost when using a distribution-free test. Therefore standard tests are applied even when the conditions for their strict validity are not satisfied. Very often, the information "lost" never existed, but there may be some information regarding the distribution, e.g. that it is symmetric, unimodal, etc. This information is usually lost when using a distribution-free test, since most of them are distribution-free with respect to very general classes of distributions. To increase the applicability of distribution-free tests, it seems important to find optimal tests with respect to more restricted classes Ω, or in any case to rank given tests with respect to such classes.

For the comparison of tests, the minimax principle seems very well fitted. The criticism against it (cf. p. 138) may to some extent be taken into account. If HOEFFDING's method of arranging the elements of ω_2 according to their distance from ω_1 is used, no numerical weights have to be given to them. If moreover, Ω is restricted to include only those distributions that are relevant, using all a priori information, the method of looking primarily to the least favorable of these cannot be criticized as much as when Ω is too wide. The following section is devoted to the development of the theoretical set-up necessary for a comparison based on these principles.

5. The new criterion

In consequence with the minimax principle, we may define a test φ' as being *better in the minimax sense* than another test φ with respect to a given weight function, if

$$\sup_{F\in\Omega} r(\varphi', F) < \sup_{F\in\Omega} r(\varphi, F).$$

Of course, the best test according to this criterion is identical with the minimax test defined on p. 138.

Specifically, we may compare all tests $\varphi\in\Phi$ and find the best one among these in the minimax sense. Let us call it $\varphi^{(m)}$. Then, by definition,

$$\sup_{F\in\Omega} r(\varphi^{(m)}, F) = \min_{\varphi\in\Phi} \sup_{F\in\Omega} r(\varphi, F).$$

Different weight functions will usually yield different $\varphi^{(m)}$. If for two families Φ' and Φ the $\varphi'^{(m)}$ relative to any weight function belonging to a given class is better than the corresponding $\varphi^{(m)}$, Φ' may be said to be *superior* to Φ in the minimax sense with respect to that class.

As in the case of optimal tests, we may take regard to r_1 only, and compare tests of the same size. Thus, φ'_α may be said to be *conditionally better* than φ_α in the minmax sense if

$$\sup_{F\in\omega_2} r_1 (\varphi'_\alpha, (F) < \sup_{F\in\omega_2} r_1 (\varphi_\alpha, F).$$

Now, by an obvious extension of HOEFFDING's (1951) theorem 1 on optimal tests, we have the following

Theorem 2. If for two tests of size α,

$$\inf_{F\in\Omega(a)} P(d_2 \mid \varphi'_\alpha, F) > \inf_{F\in\Omega(a)} P(d_2 \mid \varphi_\alpha, F) \text{ for all } a, \tag{2}$$

then φ'_α is conditionally better than φ_α in the minimax sense with respect to any a-ordered incomplete weight function.

The proof is analogous to HOEFFDING's: If (2) is true, then

$$\sup_a W_a \sup_{F\in\Omega(a)} P(d_1 \mid \varphi'_\alpha, F) < \sup_a W_a \sup_{F\in\Omega(a)} P(d_1 \mid \varphi_\alpha, F) \tag{3}$$

where W_a is the constant value of $W(F, d_1)$ for all $F\in\Omega(a)$.

(3) may also be written

$$\sup_{a} \sup_{F\in\Omega(a)} W(F, d_1)\, P(d_1 \mid \varphi'_\alpha, F) < \sup_{a} \sup_{F\in\Omega(a)} W(F, d_1)\, P(d_1 \mid \varphi_\alpha, F)$$

which is by definition equivalent to

$$\sup_{F\in\omega_2} r_1(\varphi'_\alpha, F) < \sup_{F\in\omega_2} r_1(\varphi_\alpha, F)$$

and φ'_α is conditionally better than φ_α in the minimax sense.

Now, if for all α the tests of a family Φ' are conditionally better than those of another family Φ with respect to all a-ordered incomplete weight functions, then Φ' may be said to be *conditionally superior* to Φ in the minimax sense with respect to $\{\Omega(a)\}$.

A weight function where $W(F, d_1)$ is a-ordered and where

$$W(F, d_2) = \begin{cases} V & \text{for all } F \in \omega_1 \\ 0 & \text{for all } F \in \omega_2 \end{cases}$$

where V is a constant, may be called an *a-ordered complete* weight function. Using such a weight function, we have for all tests of distribution-free families

$$r_2(\varphi_\alpha, F) = P(d_2 \mid \varphi_\alpha, F)\, W(F, d_2) = \alpha V \text{ for all } F \in \omega_1.$$

Further $r = r_1 + r_2$, and $r_i(\varphi, F) = 0$ for $F \notin \omega_i$. Thus, if Φ' is conditionally superior to Φ with respect to the corresponding a-ordered incomplete weight function, it follows that

$$\sup_{F\in\Omega} r(\varphi'_\alpha, F) \le \sup_{F\in\Omega} r(\varphi_\alpha, F) \text{ for any } \alpha,$$

i.e. φ'_α is at least as good as φ_α in the minimax sense. Now, let $\varphi^{(m)}$ be of size α_0. Then φ'_{α_0} must be at least as good as $\varphi^{(m)}$, and a fortiori $\varphi'^{(m)}$ is also at least as good as $\varphi^{(m)}$ for any a-ordered complete weight function. Thus, we have established

Theorem 3. If two test families Φ' and Φ are distribution-free with respect to ω_1, and if Φ' is conditionally superior to (or equivalent to) Φ in the minimax sense with respect to $\{\Omega(a)\}$, then Φ' is never inferior to Φ in the minimax sense with respect to the class of a-ordered complete weight functions.

As a corollary to Theorems 2 and 3 we may state:

A test family Φ' of maximin power with respect to $\{\Omega(a)\}$ is conditionally superior to any other test family, with the exception of other test families of maximin power which, if they exist, are equivalent to Φ' in the minimax sense, with respect to any a-ordered incomplete weight function. If Φ' is distribution-free, it is furthermore superior (or equivalent) to any other distribution-free test family, but not necessarily to any other test family, with respect to any a-ordered complete weight function.

The consequence of the theorems is that for distribution-free test families, if (2) is satisfied for all α, Φ' is better than Φ according to rather general optimality criteria, whether it is adequate to compare the two kinds of wrong decisions or not. For test families that are not distribution-free, the relation between the tests can be assured to hold only for each size α separately, but even so, this way of comparing two tests seems attractive. Thus we may very often reduce the problem of comparing two test families to the following steps:

a) fix ω_1 and ω_2,

b) find a partition $\{\Omega(a)\}$ of ω_2 such that F_1 may be said to be further away from ω_1 than F_2 if $F_1 \in \Omega(a_1)$, $F_2 \in \Omega(a_2)$, and $a_1 > a_2$,

c) for arbitrary fixed a and α, find out which test has the bigger lower limit of its power function, and

d) see if this relation is independent of a and α.

It may, of course, happen that one test is better for some a or α, and worse for others. In that case, the weight function has to be specified numerically in order to get a discrimination between the tests, and the use of the minimax principle may be more severely criticized. In the following sections, however, we shall be able to show that there are interesting cases where the relation between the tests is independent of a and α, and thus a fairly unambiguous choice between them can be made.

Before going over to the examples it is important to notice that the properties of superiority and conditional superiority depend rather heavily on the choice of Ω. If Ω is very wide, the test family Φ' may be superior to Φ, but if Ω is narrowed down, corresponding to an increase of a priori information, Φ may well become superior to Φ'. If this is the case, it would be of interest to find out when this change takes place.

6. Location tests

In the remaining part of the paper, the general principles, given in the previous section, of comparing two test families, are applied to the problem of testing the value of the median μ in a one-dimensional distribution. The discussion is confined to the one-sided case of testing $\mu = 0$ against $\mu > 0$. The sample is assumed to consist of n independent observations drawn from the population to be tested.

In order to be able to introduce successively increasing information (by decreasing Ω) symmetrically into ω_1 and ω_2, we shall use the following notation:

Let F	denote the distribution of the one-dimensional variable, and
$F(X)$	the cumulative distribution function belonging to it,
M_1	the class of all one-dimensional distributions with $\mu = 0$,
M_2	the class of all one-dimensional distributions with $\mu > 0$,
Ω'	a class of one-dimensional distributions being considered, e.g. distributions with symmetrical density functions,
ω_1	$= M_1 \cap \Omega'$
ω_2	$= M_2 \cap \Omega'$
Ω	$= \omega_1 \cup \omega_2$.

By different choices of Ω' the problem may be given different degrees of generality. If Ω' contains only normal populations, it is the classical Student problem of testing the mean of a normal population with unknown variance. For changing Ω', or equivalently, changing degrees of information, the order between the tests according to our optimality criterion will probably change. We shall be concerned with a comparison of tests at a gradual decrease of Ω' from the most general case to more specific ones.

Like Hoeffding, we shall define a partition of ω_2, and say that of two distributions in ω_2, that one is farther away from ω_1 which has the bigger $a = \frac{1}{2} - F(0)$. Besides being mathematically simple, this seems to be a relevant measure in several kinds of applications. $F(0)$ will be denoted by q. For convenience, we shall write $\omega_2(q)$ for $\Omega(\frac{1}{2} - q)$, and q-ordered for $(\frac{1}{2} - q)$-ordered weight functions.

In the following comparison of tests, rank tests will play a predominant part, and it seems appropriate first to examine this type of tests.

7. Rank test families

A rank test family may be defined in the following way. Let the sample point be

$$x = \{x_i\} = \{\varepsilon_i y_i\} \ (i = 1, 2, \ldots, n),$$

where $y_i = |x_i|$ and $\varepsilon_i = \operatorname{sgn} x_i$. We shall assume that the distributions are continuous, so that $P\{x_i = 0\} = 0$, and $P\{y_i = y_j\} = 0$ for $i \neq j$. The observations may then be ordered in such a way that

$$y_1 < y_2 < \cdots < y_n.$$

If $\psi(x)$ is a function of $\{\varepsilon_i\}$ only, the fixed limit test family and the permutation test family generated by $\psi(x)$ coincide and constitute a rank test family. Any sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ defines a region in the sample space. There are $M = 2^n$ such *basic cells* (Walsh, 1949). Let us denote them by $B_j (j = 1, \ldots, M)$.

For distributions with density functions symmetrical around 0,

$$P\{x \in B_j \mid F \in \omega_1\} = \frac{1}{M} \quad \text{for any } j.$$

Thus, all rank test families are distribution-free in the class of such symmetrical distributions.

Any ψ defines an ordering of the B_j, say $B_{t_1}, B_{t_2}, \ldots, B_{t_M}$, in the sense that $\psi(B_{t_1}) \geq \psi(B_{t_2}) \geq \cdots \geq \psi(B_{t_M})$. A rank test family is completely defined by the sequence $t_1, t_2, \ldots, t_M$. Thus, any strictly increasing function of ψ will generate the same rank test family as ψ.

An alternative way of describing certain rank test families was pointed out by Walsh (1949), who arranged the observations in increasing order of magnitude. Let us call the i^{th} observation in that order $x_{(i)}$, so that

$$x_{(1)} < x_{(2)} < \ldots < x_{(n)}.$$

Then he suggested the tests

$$\varphi(x) = \begin{cases} 1 & \text{if } \min\left[\frac{1}{2}(x_{(i)} + x_{(j)}),\ \frac{1}{2}(x_{(k)} + x_{(l)}), \ldots\right] > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $i, j, k, l, \ldots$ are determined so as to give the desired size to the test. Walsh showed that an increase in one or several of the $i, j, k, l, \ldots$ amounts to increasing the region where $\varphi(x) = 1$ by one or more basic cells. This means, in the present notation, that to each test family of WALSH's type corresponds an ordering of the B_j, or equivalently, a function $\psi(x)$. The opposite is not true, at is easily seen by an example: Take for simplicity $n=5$. Then, if $\psi(x)=\sum_{i=1}^{5}\varepsilon_i\, i$, for $c=4$ we get $P\{\psi(x)=c\}=0$ and $\alpha=\frac{10}{32}$. Here, $\psi(+--++)=$ $=\psi(++++-)=5$. In order to get $\varphi(x)=1$ for these sequences also in WALSH's formulation, we can at most require $\min\,[\frac{1}{2}(x_{(2)}+x_{(4)}),\, x_{(3)}]>0$ but that is also satisfied by the sequences $(-+++-)$, and $(+-+-+)$ with $\psi(x)=3$, and by $(+-++-)$ with $\psi(x)=1$. WALSH's formulation of the tests is, however, very simple in numerical applications and could profitably by used whenever possible.

The most powerful rank test family for the case when ω_2 contains one distribution F' only is generated by the function

$$\psi(B_j)=P\{x\in B_j\,|\,F'\}$$

(cf. LEHMANN and STEIN, 1949, and HOEFFDING, 1951).

When working exclusively with rank tests we may introduce slightly modified admissibility concepts. Thus, a test is *α-admissible among rank tests* if no uniformly more powerful rank test exists, and it is *admissible among rank tests* if no uniformly better rank test exists. These two concepts are, however, equivalent. It is trivial that admissibility among rank tests implies α-admissibility among rank tests, and the converse is proved by

Theorem 4. A rank test that is α-admissible among rank tests is also admissible among rank tests for any weight function.

Proof: Let φ'_{α_1} be α-admissible among rank tests. Now suppose the theorem were not true. Then there would exist a rank test φ_{α_2} that is uniformly better than φ'_{α_1}.

a) Suppose first that $\alpha_2>\alpha_1$. But then

$$P(d_2\,|\,\varphi_{\alpha_2},F)>P(d_2\,|\,\varphi'_{\alpha_1},F)\quad\text{for all}\quad F\in\omega_1$$

and φ_{α_2} cannot be uniformly better than φ'_{α_1}.

b) If $\alpha_1=\alpha_2$,

$$P(d_2\,|\,\varphi_{\alpha_1},F)\le P(d_2\,|\,\varphi'_{\alpha_1},F)\quad\text{for all}\quad F\in\omega_1,$$

and

$$P(d_2\,|\,\varphi_{\alpha_1},F)\ge P(d_2\,|\,\varphi'_{\alpha_1},F)\quad\text{for all}\quad F\in\omega_2,$$

with the inequality sign true for at least one $F\in\Omega$. But since the tests are similar and of the same size,

$$P(d_2\,|\,\varphi_{\alpha_1},F)=P(d_2\,|\,\varphi'_{\alpha_1},F)=\alpha_1,\ \text{for all}\ F\in\omega_1.$$

Thus, the inequality sign must hold for some $F\in\omega_2$. But that would imply that φ_{α_1} is uniformly more powerful than φ'_{α_1}, which is impossible, since φ'_{α_1} is α-admissible among rank tests.

c) If $\alpha_2 < \alpha_1$, we may compare φ_{α_2} with the test of size α_1 from the same family. As the power with respect to any absolutely continuous $F \in \omega_2$ is an increasing function of α,

$$P(d_2 \mid \varphi_{\alpha_2}, F) < P(d_2 \mid \varphi_{\alpha_1}, F) \quad \text{for all } F \in \omega_2.$$

Thus, if

$$P(d_2 \mid \varphi_{\alpha_2}, F) \geq P(d_2 \mid \varphi'_{\alpha_1}, F) \quad \text{for all } F \in \omega_2$$

φ_{α_1} would be uniformly more powerful than φ'_{α_1}, which is impossible by definition. This completes the proof.

A family of rank tests, all members of which are admissible among rank tests, may be called *rank-admissible*. Although we may perhaps, because of its simplicity, accept a rank test family that is not admissible, we would hardly do so, were it not rank-admissible.

For later reference, we list a few rank test families below:

1. The sign test family, apparently first mentioned by FISHER (1925), and investigated by DIXON and MOOD (1946). It is generated by $\psi(x) = \sum_{i=1}^{n} \varepsilon_i$.

In a sense, the sign tests are quite crude, as ψ has the same value for several B_j. It will be convenient in this connection to introduce another notation. Let T_k be the totality of all B_j with $\psi(B_j) = k$. To point out which $B_j \in T_k$, they will occasionally be numbered $B_{k1}, B_{k2}, \ldots, B_{km_k}$. Since $P(T_k \mid \omega_1)$ will be rather big, the tests depend heavily on the outcome of the random experiment determining the decision when $\varphi(x) = p$. HEMELRIJK (1950 a, b) pointed out, however, that a sign test may be regarded as basic to every other test in this situation. Thus, any test for the median of a symmetrical distribution may be considered as a combination of a sign test and a conditonal two-sample test, testing whether or not the positive and the negative observations come from identical populations. The functions generating these tests may be called $\psi_1(x)$ for the sign tests and $\psi_2(x)$ for the two-sample tests. Then, the combined test family is generated by a function $\psi(\psi_1, \psi_2)$. HEMELRIJK suggested a rule for the construction of this function, but gave no justification for it. However, in analyzing different tests, it may sometimes be convenient to discuss separately the choice of ψ_2 and the construction of ψ. As an example of this, we shall describe below three test families which are based on the same ψ_2, but where the construction of ψ is different. The common ψ_2 is $\psi_2 = \sum_{i=1}^{n} \varepsilon_i \, i$, and the test families are:

2a. Wilcoxon's one-sample test family, proposed by WILCOXON (1945). It has $\psi = \psi_2$. This test may be regarded as the rank analogue of the permutation test family generated by $\psi(x) = \bar{x}$, suggested by FISHER (1935).

2b. The Wilcoxon-Hemelrijk test family. The principle of constructing ψ given by HEMELRIJK is easily expressed in the present notation. Let B_{ki} be numbered so that for all k

$$\psi_2(B_{k1}) \geq \psi_2(B_{k2}) \geq \cdots \geq \psi_2(B_{k m_k}).$$

Then, for $k = \psi_1(B_{ki}) > 0$,

$$\psi(B_{ki}) = \begin{cases} 1/i & \text{if } \psi_2(B_{ki}) > \psi_2(B_{k(i+1)}) \\ 1/(i+r) & \text{if } \psi_2(B_{ki}) = \psi_2(B_{k(i+1)}) = \cdots = \psi_2(B_{k(i+r)}) > \psi_2(B_{k(i+r+1)}). \end{cases}$$

For $k \leq 0$, ψ is not defined.

2c. The Wilcoxon-sign test family. Here, $\psi = n^2\psi_1 + \psi_2$ which means that $\psi(B_{ki}) > \psi(B_{lj})$ if $k > l$, and the value of ψ_2 is of relevance if $k = l$ only. This is a sign test family, modified so as to make the influence of the random experiment smaller.

To make the difference between test families 2a–c clear, we give an example of the orders, defined by the different ψ, of the B_j when $n = 5$.

				Order among B_j		
		$\psi_1 = \Sigma \varepsilon_i$	$\psi_2 = \Sigma \varepsilon_i i$	2a	2b	2c
T_5:	B_{51} + + + + +	5	15	1	1	1
T_3:	B_{31} − + + + +	3	13	2	2, 3	2
	B_{32} + − + + +	3	11	3	4, 5	3
	B_{33} + + − + +	3	9	4, 5	6	4
	B_{34} + + + − +	3	7	6, 7	7, 8, 9	5
	B_{35} + + + + −	3	5	8, 9, 10	10	6
T_1:	B_{11} − − + + +	1	9	4, 5	2, 3	7
	B_{12} − + − + +	1	7	6, 7	4, 5	8
	B_{13} − + + − +	1	5	8, 9, 10	7, 8, 9	9, 10
	B_{14} + − − + +	1	5	8, 9, 10	7, 8, 9	9, 10
etc.						

Given a rank test family, F' is a least favorable distribution in $\omega_2(q)$ relative to it, if

$$\sum_{i=1}^{s} P(B_{t_i} \mid F') = \min_{F\in\omega_2(q)} \sum_{i=1}^{s} P(B_{t_i} \mid F) \quad \text{for all } s = 1, 2, \dots, M.$$

Thus, $P(B_j \mid F')$ should, as far as possible, be a decreasing function of $\psi(B_j)$. The value of the minimum power depends on the restrictions expressed by the size of Ω'. Under all circumstances, however, for all $F \in \omega_2(q)$, and for all $B_{ki} \in T_k$, $\sum_{i=1}^{m_k} P(B_{ki} \mid F) = q^{\frac{n-k}{2}}(1-q)^{\frac{n+k}{2}}$. This means that if for a rank test family generated by a function $\psi(\psi_1, \psi_2)$ with $\psi_2(B_{k1}) \geq \psi_2(B_{k2}) \geq \cdots \geq \psi_2(B_{km_k})$, $\sum_{i=1}^{s} P(B_{ki} \mid F)$ is minimized by the same distribution F' for all s and k, then F' is a least favorable distribution in $\omega_2(q)$ relative to all tests with the same ψ_2, irrespective of the form of ψ. Thus, the search for a least favorable distribution can most profitably start by considering each T_k separately.

As the rank tests make no use of the absolute magnitudes of the y's, any topological transformation of the y scale does not affect the test. By choosing a suitable transformation the calculations may be made easier. For instance, the positive side of the distribution may be made rectangular over the range $(0, 1-q)$. The same transformation is then applied to the absolute values of the negative side. As in this way the positive side is made identical for all distributions with common q, only the negative side will be of relevance in the following discussion. The transformation is defined by

$$x' = [F(|x|) - q] \operatorname{sgn} x.$$

The transformed cumulative distribution function becomes

$$F_T(x') = \begin{cases} 0 & \text{for } x' < q-1 \\ F[-F^{-1}(q-x')] & q-1 \le x' < 0 \\ x'+q & 0 \le x' < 1-q \\ 1 & x' \ge 1-q. \end{cases}$$

The density function is most easily expressed in terms of x:

$$f_T(x') = \begin{cases} 0 & \text{for } x' < q-1 \\ -\dfrac{d\,F(x)}{d\,F(-x)} = \dfrac{f(x)}{f(-x)} & q-1 \le x' < 0 \\ 1 & 0 \le x' < 1-q \\ 0 & x' \ge 1-q. \end{cases}$$

It should be noted that for $q=\frac{1}{2}$, i.e. $F \in \omega_1$, F_T is for all symmetrical distributions rectangular $(-\frac{1}{2}, \frac{1}{2})$. Thus, if Ω' contains only symmetrical distributions, ω_1 is reduced to one element.

As an example of the effect of the transformation for $F \in \omega_2$, let us regard the Cauchy distribution with $F(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arc\,tg}(x-a)$ where $q = \frac{1}{2} - \frac{1}{\pi} \operatorname{arc\,tg} a$, and $F \in \omega_2(q)$. For $q-1 \le x' < 0$, we get

$$F^{-1}(q-x') = a + \operatorname{tg}(q - x' - \tfrac{1}{2})$$

$$F_T(x') = \tfrac{1}{2} - \frac{1}{\pi} \operatorname{arc\,tg}\,(2a + \operatorname{tg} \pi (q - x' - \tfrac{1}{2}))$$

and, after some calculation,

$$f_T(x') = [1 + 2a(a - a\cos 2\pi x' - \sin 2\pi x')]^{-1}.$$

This function has a minimum for $x' = \dfrac{q-1}{2}$, and is symmetrical around this point with $f_T(q-1) = f_T(0) = 1$.

For the normal distribution, the explicit expression for $F_T(x')$ is difficult to obtain, but by expressing $f_T(x')$ in terms of x, it is possible to obtain its value in some points. Thus, for $N(\frac{1}{2}, \sigma^2)$, where σ^2 has to be determined so as to make $F(0) = q$,

$$f_T(x') = \exp\left\{\frac{x}{\sigma^2}\right\} \qquad \text{for } q-1 \le x' < 0.$$

As $x' = q-1$ corresponds to $x = -\infty$ and $x' = 0$ to $x = 0$, we get $f_T(q-1) = 0$ and $f_T(0) = 1$.

We can also express $\dfrac{d\,f_T(x')}{d\,x'}$ in terms of x, and get

$$\frac{d\,f_T(x')}{d\,x'} = \text{constant} \cdot \exp\left\{\frac{1}{2\sigma^2}(x^2 + 3x + \tfrac{1}{4})\right\} \qquad \text{for } q-1 \le x' < 0.$$

Thus, for $x' = q - 1$, or $x = -\infty$, the derivative is infinite. As x and x' increase to 0, it is first decreasing, then increasing, but is always positive.

Generally, for any distribution with symmetrical and unimodal density function, $f(x)/f(-x)$ can never exceed 1 for $x < 0$, as long as $q \leq \frac{1}{2}$, and accordingly $f_T(x') \leq 1$ for $x' < 0$.

A large group of transformed distributions have symmetrical equivalents in Ω', as is shown by

Theorem 5. To every $J(x')$ satisfying

$$J(x') \begin{cases} = 0 & \text{for } x' \leq q-1 \\ \left.\begin{array}{l}\text{strictly increasing} \\ < x'+1-q\end{array}\right\} & \text{for } q-1 < x' < 0 \\ = x'+q & \text{for } 0 \leq x' < 1-q \\ = 1 & \text{for } x' \geq 1-q \end{cases} \tag{4}$$

there corresponds at least one symmetrical distribution for which $F_T(x') = J(x')$.

Proof: Let F have $\mu = \frac{1}{2}$. Then, according to the symmetry,

$$F(x) = 1 - F(1-x). \tag{5}$$

Now, take in the interval $(0, \frac{1}{2})$ an arbitrary increasing function $F_1(x)$, satisfying

$$F_1(0) = q \text{ and } F_1(\tfrac{1}{2}) = \tfrac{1}{2}.$$

Then, define $F(x)$ successively in the following way:

$$F(x) = \begin{cases} F_1(x) & \text{for } \quad 0 \leq x < \frac{1}{2} \\ 1 - F_1(1-x) & \frac{1}{2} \leq x < 1 \\ J[-F_1(-x)+q] & -\frac{1}{2} \leq x < 0 \\ J[q-1+F_1(1+x)] & -1 \leq x < -\frac{1}{2} \\ 1 - J[q - F_1(x-1)] & 1 \leq x < 1\frac{1}{2} \\ 1 - J[q-1+F_1(2-x)] & 1\frac{1}{2} \leq x < 2 \\ \text{etc.} & \end{cases}$$

using the recursive formula

$$F(x) = \begin{cases} J[q - F(-x)] & \text{for } x < 0 \\ 1 - F(1-x) & \text{for } x > 0. \end{cases} \tag{6}$$

By construction, $F(x)$ satisfies the symmetry condition (5). It remains to show that $F(x)$ is a cumulative distribution function, i.e. that it is non-decreasing, and that $F(-\infty) = 0$, and $F(\infty) = 1$.

In the recursive formula (6), it is seen that $F(x)$ is increasing within any interval of the left-hand side if it is increasing in the corresponding interval of the right-hand side. Thus, since $F_1(x)$ is increasing, $F(x)$ is increasing within any interval where the formula is applied. But we must also demand that the value of $F(x)$ for $x > 0$ within one interval $(i, i+1)$ shall be greater than that obtained in the previous interval $(i-1, i)$, thus

$$F(x) > F(x-1) \quad \text{for } i \le x \le i+1, \quad i = 1, 2, \ldots \tag{7}$$

But according to (6),

$$F(x) = 1 - F(1-x) = 1 - J[q - F(x-1)].$$

Put $x' = q - F(x-1)$. Then (7) becomes

$$1 - J(x') > q - x' \quad \text{for } x' < 0$$

$$\text{or} \quad J(x') < x' + 1 - q.$$

This condition is satisfied for $q - 1 < x' < 0$ according to (4), and $F(x)$ is increasing in the whole of the interval $0 < F(x) < 1$. Thus, $F(x)$ is a cumulative distribution function.

A consequence of theorem 5 is that to any class Ω'_T of F_T with cumulative distribution functions satisfying (4), there corresponds a class Ω' of symmetrical distributions, and to a certain extent we may in the following discussion consider the classes Ω' and Ω'_T interchangeably.

8. Comparison of location tests

A. We are now in a position to make the comparison between test families outlined in section 6 above. For Ω' = the class of all absolutely continuous distributions, HOEFFDING (1951) proved that the sign test family is of maximin power with respect to $\{\omega_2(q)\}$. Thus, according to the corollary to our Theorems 2 and 3, it is conditionally superior to any other test family and superior to any other rank test family with respect to q-ordered weight functions, in both cases with the possible exception of other test families of maximin power. However, if there exists another family of maximin power, its tests may be uniformly more powerful and uniformly better than the sign tests, in which case the sign test family will not be admissible or even rank-admissible. HOEFFDING did not discuss this possibility. We shall study it in subsections B and E below.

B. Let us now reduce Ω' to distributions with symmetrical density functions and call this class Ω'_B. Then the mean of the distribution is, if it exists, equal to the median, and the tests may be considered as tests of the mean. We shall see that the situation is not changed from A above. For let $F \in \omega_1$ and $G \in \omega_2(q)$ be two distributions with density functions

$$f(x) = \frac{1-2q}{2(1-q)} \left(\frac{q}{1-q}\right)^{[|x|]}$$

$$g(x) = (1-2q) \left(\frac{q}{1-q}\right)^{|[x]|}$$

where $[x]$ means the largest integer $\leq x$, also for $x < 0$. Thus,

$$|[x]| = [|x|] \text{ for } x > 0, \text{ and for all integer } x, \text{ and}$$

$$|[x]| = [|x|] + 1 \text{ for non-integer } x < 0.$$

Now, since $P\{x = \text{an integer}\} = 0$, we get

$$\prod_{i=1}^{n} \frac{g(x_i)}{f(x_i)} = 2^n q^{n_-} (1-q)^{n_+}$$

where n_- and n_+ are the number of negative and positive among the n observations. For fixed n and q, this is obviously a function of n_+ only. By Neyman-Pearson's lemma, a test family generated by $\psi(x) = n_+$, or, equivalently, by $\psi(x) = n_+ - n_-$, is most powerful for the case when $\omega_1 = F$ and $\omega_2(q) = G$. This family is the sign test family. As a special case of HOEFFDING's (1951) theorems 1 and 2, a similar test φ'_α is of minimax risk r_1 with respect to any q-ordered incomplete weight function, if for any fixed q it is most powerful for a case where $\omega_1 = F'$ and $\omega_2(q) = F''_q$, and

$$P(d_2 \mid \varphi'_\alpha, F) \geq P(d_2 \mid \varphi'_\alpha, F''_q) \quad \text{for all } F \in \omega_2(q).$$

In other words, F''_q should be a least favorable distribution in $\omega_2(q)$ to φ'_α. As the equality sign holds true for all tests of the sign test family, it is, according to our previous results, conditionally superior or equivalent (since there may be more than one test family of minimax risk) to any other test family and superior or equivalent to any other rank test family, with respect to any q-ordered weight function.

The transform of G, as well as of the distribution used by HOEFFDING (1951, p. 91) to demonstrate the optimality of the sign test family, has the density function $f_T(x') = \frac{q}{1-q}$ for $q - 1 \leq x' < 0$. For reference purposes it may be convenient to give a name to this distribution, and we are going to call it the *bi-rectangular* distribution. It gives equal probabilities to all sequences B_{ki} belonging to one T_k. Thus, in the case when ω_2 contains this distribution only, all tests with $\psi = k\psi_1 + \psi_2$, where $k > \max|\psi_2|$, are equivalent and most powerful rank tests. Let us call test families generated by such functions *modified sign test families*. The tests belonging to such a family are essentially sign tests, but the random order between the B_{ki} within a T_k that is characteristic for the sign test family is replaced by a more or less complete order. The Wilcoxon-sign test family, described above, is an example of a modified sign test family.

We shall now state two important theorems on modified sign test families. These theorems will be fundamental to the following discussion.

Theorem 6. If the bi-rectangular distribution belongs to $\omega_2(q)$, no rank test family that is not a modified sign test family can be superior in the minimax sense to the sign test family with respect to all q-ordered weight functions.

Theorem 7. The sign test family is not rank-admissible if there exists a modified sign test family, the tests of which are uniformly more powerful than those of the sign test family. Such modified sign test families are all equally good in the minimax sense and also as good as the sign test family, but superior to any other rank test family.

Proof of Theorem 6: The tests belonging to a family that is not a modified sign test family must be uniformly less powerful than (or exceptionally as powerful as) those of the sign test family against the bi-rectangular distribution. Since the sign tests have the same power against all distributions in $\omega_2(q)$, the theorem follows.

Theorem 7 is true by definition.

It seems intuitively reasonable to assume that the sign test family is rank-admissible when Ω' is wide, say at least equal to Ω'_B. This has not yet been proved, but no common test family is superior to the sign test family. No modified sign test families seem to have been suggested in the literature. However, if we construct such families by using for ψ_2 the ψ of the most common one-sided rank tests, the sign test family will be superior to all of them. This is so because all such families, including the Wilcoxon-sign test family, satisfy

$$\psi_2(\varepsilon_1, \varepsilon_2, \cdots + - \cdots, \varepsilon_n) \leq \psi_2(\varepsilon_1, \varepsilon_2, \cdots - + \cdots, \varepsilon_n)$$

where all ε_i are identical in the two sequences except that a plus sign and a neighboring minus sign have changed places. A distribution satisfying

$$\frac{d f_T(x')}{d x'} < 0 \qquad \text{for } q-1 < x' < 0$$

will give high probability to the basic cells with small ψ_2 and vice versa. Thus, the tests will for almost all α be less powerful than the corresponding sign tests against that distribution. In Ω'_B therefore, no test family has so far been proved to have tests that are uniformly more powerful than those of the sign test family.

C. As the proofs above are still true, if $f_T(x') \leq 1$ for $q-1 \leq x' < 0$, nothing in the above discussion is changed when Ω' is reduced to unimodal, symmetrical distributions.

D. When further reducing Ω', it may seem natural to do away with some pathological distributions by asking that

$$\nu = P\{x_1 > -x_2 \mid x_1 > 0, x_2 < 0\} \geq \tfrac{1}{2}.$$

The class of distributions satisfying this condition will be called Ω'_D. The condition may also be written $P(-\,+) \geq P(+\,-)$, since

$$\nu = \frac{P(-\,+)}{P(-\,+) + P(+\,-)}.$$

For ω_1, ν is always $\frac{1}{2}$ according to the symmetry, and the reduction is effective in ω_2 only. In the transform, the condition may be written

$$\frac{1}{q(1-q)} \int_{q-1}^{0} (1-q+x')\, d F_T(x') \geq \tfrac{1}{2},$$

or

$$\frac{1}{q} \int_{q-1}^{0} x'\, d F_T(x') \geq -\frac{1-q}{2}.$$

This condition implies that the mean of the negative side of the transformed distribution must not be situated to the left of the midpoint of the interval $(q-1, 0)$ which is its range. As the transformed density function of Cauchy's distribution and of the bi-rectangular distribution are both symmetrical around $\frac{q-1}{2}$, it is clear that $\nu = \frac{1}{2}$ for these distributions. Thus, they belong to the class but are border-line cases.

As the bi-rectangular distribution belongs to Ω'_D, the sign test family is of maximin power. But it is not rank-admissible for all n and α in this class. For $P(-+) \geq P(+-)$. If $n=2$, the ψ_1-values of the two sequences are equal, and any test of size $\alpha = \frac{1}{2}$ belonging to a modified sign test family, satisfying $\psi_2(-+) > \psi_2(+-)$, is uniformly more powerful than the sign test of the same size.

E. Next, define Ω'_E as the class of all distributions, the transformed densities of which are non-decreasing in $(q-1, 0)$. This implies that the mean of the negative side is at least $\frac{q-1}{2}$, and it follows that Ω'_E is a subclass of Ω'_D. It is also effectively a subclass, since obviously the Cauchy distribution does not belong to Ω'_E.

We shall prove that in Ω'_E

$$P(\varepsilon_1, \varepsilon_2, \cdots + - \cdots, \varepsilon_n) \leq P(\varepsilon_1, \varepsilon_2, \cdots - + \cdots, \varepsilon_n)$$

where all ε_i are identical in the two sequences except that a plus sign and a neighboring minus sign have changed places. Denote by s_r the y-number of the r^{th} negative observation, i.e. of $x_{(n_-+1-i)}$. Let the y-number of the specified negative sign be $s_j = i+1$ in the left hand side of (8) and $s_j = i$ in the right hand side. We shall assume that $f_T(x')$ is twice continuously differentiable, which means no essential restriction to the class. We may write the two sides of (8) in the form of multiple integrals

$$\int_{q-1}^{0}\int_{q-1}^{x'_1}\cdots\int_{q-1}^{x'_{n-1}} h_1(x'_1, \ldots, x'_{i-1})\, f_T(x'_{i+1})\, h_2(x'_{i+2}, \ldots, x'_n)\; d\,x'_1 \ldots d\,x'_n \qquad \text{(L.H.S.)}$$

and

$$\int_{q-1}^{0}\int_{q-1}^{x'_1}\cdots\int_{q-1}^{x'_{n-1}} h_1(x'_1, \ldots, x'_{i-1})\, f_T(x'_i)\, h_2(x'_{i+2}, \ldots, x'_n)\, d\,x'_1 \ldots d\,x'_n \qquad \text{(R.H.S.)}$$

where

$$h_1 = \prod_{k=1}^{j-1} f_T(x'_{s_k}) \qquad\qquad h_2 = \prod_{k=j+1}^{n_-} f_T(x'_{s_k}).$$

If the relation (8) is to hold for any combination of signs

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{i-1}, \varepsilon_{i+2}, \ldots, \varepsilon_n,$$

i.e. for any h_1 and h_2, the following must hold identically in x'_{i-1}:

$$\int_{q-1}^{x'_{i-1}}\int_{q-1}^{x'_i} f_T(x'_{i+1})\, H(x'_{i+1})\, d\,x'_i\, d\,x'_{i+1} \leq \int_{q-1}^{x'_{i-1}}\int_{q-1}^{x'_i} f_T(x'_i)\, H(x'_{i+1})\, d\,x'_i\, d\,x'_{i+1}$$

where

$$H(x'_{i+1}) = \int\limits_{q-1}^{x'_{i+1}} \cdots \int\limits_{q-1}^{x'_{n-1}} h_2(x'_{i+2}, \ldots, x'_n)\, d x'_{i+2} \ldots d x'_n.$$

This is clearly fulfilled if for all x'_i

$$\int\limits_{q-1}^{x'_i} f_T(x'_{i+1})\, H(x'_{i+1})\, d x'_{i+1} \leq f_T(x'_i) \int\limits_{q-1}^{x'_i} H(x'_{i+1})\, d x'_{i+1}$$

and, as both sides are $=0$ for $x'_i = q-1$, this is in turn true if the derivatives with respect to x'_i fulfill the same relation for all x'_i:

$$f_T(x'_i)\, H(x'_i) \leq \frac{d f_T(x'_i)}{d x'_i} \int\limits_{q-1}^{x'_i} H(x'_{i+1})\, d x'_{i+1} + f_T(x'_i)\, H(x'_i).$$

As $H(x) \geq 0$, this means that

$$\frac{d f_T(x'_i)}{d x'_i} \geq 0 \qquad \text{for } q-1 \leq x' \leq 0 \tag{9}$$

which is the definition of the class, and the relation (8) is fulfilled for all distributions belonging to it.

Ω'_E is also easy to express in terms of the original distributions. For (9) may be written

$$\frac{d}{d x} \frac{f(x)}{f(-x)} \cdot \frac{d x}{d x'} \geq 0 \qquad \text{for } x < 0.$$

As $\dfrac{d x'}{d x} = f(-x)$ for $x,\ x' < 0$, we get

$$\frac{f(x) \dfrac{d f(-x)}{d x} + f(-x) \dfrac{d f(x)}{d x}}{(f(-x))^3} \geq 0 \qquad \text{for } x < 0.$$

As $f(-x) \geq 0$, the condition becomes

$$\frac{d \log f(x)}{d x} \geq \frac{d \log f(-x)}{d x} \qquad \text{for } x < 0.$$

Now put $f_1(x) = f(x+a)$ where a is the point of symmetry. Then the condition may be written

$$\frac{d \log f_1(x-a)}{d x} \geq \frac{d \log f_1(-x-a)}{d x}.$$

This relation should hold for any $q \leq \frac{1}{2}$, i.e. for any $a \geq 0$, and for any $x < 0$. That is possible if and only if

$$\frac{d^2 \log f_1(x)}{d x^2} \leq 0$$

i.e. if and only if the logarithm of the density function is concave downwards.

To get an idea of what kind of distributions are contained in the class, let us see for some simple cases if the condition is satisfied.

For the normal distribution we get

$$f_1(x) = k e^{-c^2 x^2}$$

$$\frac{d^2 \log f_1(x)}{d x^2} = -2 c^2 < 0.$$

Thus, the normal distribution belongs to the class. For Student's distribution, we get

$$f_1(x) = c_1 \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\frac{d^2 \log f_1(x)}{d x^2} = -\frac{n+1}{2} \cdot \frac{1 - \frac{x^2}{n}}{\left(1 + \frac{x^2}{n}\right)^2}.$$

This is < 0 only when $x^2 < n$, and Student's distribution belongs to the class only when $n = \infty$, and we get the normal distribution again. Thus, the class is fairly restricted in the sense that distributions with too heavy tails are not included.

It should be noted, however, that the bi-rectangular distribution is included in Ω'_E, although on the border-line. Thus, as a consequence of Theorem 6, we need only consider modified sign test families. We shall give special attention to a certain type of such families, the tests of which will be shown to be uniformly more powerful than the corresponding sign tests.

Theorem 8. The tests of a modified sign test family with a ψ_2 satisfying

$$\psi_2(\varepsilon_1, \varepsilon_2, \ldots + - \ldots, \varepsilon_n) \leq \psi_2(\varepsilon_1, \varepsilon_2, \ldots - + \ldots, \varepsilon_n) \tag{10}$$

with the inequality sign true in some cases, are uniformly more powerful than (or exceptionally as powerful as) the corresponding sign tests.

Proof: Since the sign test family gives the same value of ψ_2 to all basic cells in a T_k, the theorem is true if it is impossible to find a distribution for which

$$\sum_{i=1}^{t} P(B_{ki}) < t\, q^{n-} (1-q)^{n+} \tag{11}$$

for at least one t. The proof will consist of four parts.

1. Let s_u denote the y-number of the u^{th} negative observation in the sequence. Suppose first that $\psi_2(B_{ki})$ is a decreasing function of s_u only, and $P(B_{ki})$ a decreasing

function of s_v only. It is no essential restriction to assume $u < v$. Denote the value of $P(B_{ki} | s_v = h)$ by p_h. The number of cells with $s_u = g$ and $s_v = h$ is

$$a_{gh} = \binom{g-1}{u-1} \binom{h-g-1}{v-u-1} \binom{n-h}{n_- - v}$$

and we want to compute the average probability of the cells with $s_u = g$:

$$\frac{\sum_{h=g+v-u}^{n-n_-+v} a_{gh} p_h}{\sum_{h=g+v-u}^{n-n_-+v} a_{gh}}.$$

The coefficient for p_h is obtained as

$$\binom{n-g}{n_- - u}^{-1} \binom{h-g-1}{v-u-1} \binom{n-h}{n_- - v}.$$

If g is increased to $(g+1)$, the coefficient for $p_h\ (h > g+v-u)$ is multiplied by

$$\frac{(h-g-v+u)\ (n-g)}{(h-g-1)\ (n-n_- - g+u)}$$

and it is seen that this ratio increases with h, or is unchanged. The coefficient for p_{g+v-u} becomes 0. Since p_h is assumed to be a decreasing function of h, this implies that the average probability of the cells with $s_u = g$ does not increase as g increases, or

$$\frac{1}{t} \sum_{i=1}^{t} P(B_{ki}) \geq q^{n_-} (1-q)^{n_+}$$

and (11) cannot be true.

2. Now, let, as before, ψ_2 be a function of s_v only, but P depend on more than one s_i, only not on s_v. If this change has any effect on the average probability, it is increasing, and (11) cannot hold.

3. Let P depend on s_u only, as in 1, and let ψ_2 depend on more than one s_i, but not on s_u. The effect is the same as in 2.

4. If P and ψ_2 depend on some common s_i, it is obvious that the power will increase, and (11) cannot hold. The theorem is proved.

If we can find a modified sign test family satisfying (10) that is also rank-admissible, this is the best we can do, as long as we restrict ourselves to rank tests. A test is clearly admissible among rank tests if it is more powerful than any other rank test for an arbitrary $F \in \omega_2$. In order to construct such a test, let us find the most powerful rank test for a series of distributions converging to the bi-rectangular distribution. Thus, define

$$f_T(x') = \frac{q\,(1+c_1)\,(1-q+x')^{c_1}}{(1-q)^{1-c_1}} \qquad \text{for } q-1 \leq x' < 0$$

where c_1 is a positive parameter. As $c_1 \to 0$, $f_T(x') \to \frac{q}{1-q}$. $f_T(x')$ is never decreasing, so the distribution belongs to Ω'_E.

Let $c_2 = \frac{q(1+c_1)}{(1-q)^{1+c_1}}$,

s_i = the y-number of the i^{th} negative observation, and

$s_0 = 0$.

The probability of a sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is equal to

$$P = n!\, c_2^{n_-} \int_0^{1-q} \int_{x_1'}^{1-q} \cdots \int_{x_{n-1}'}^{1-q} \prod_{i=1}^{n_-} (1-q-x'_{s_i})^{c_1}\, dx_1' \cdots dx_n' =$$

$$= \frac{n!\, c_2^{n_-} (1-q)^{n+c_1 n_-}}{(n-s_{n_-})! \prod_{i=1}^{n_-} \prod_{j=s_{i-1}+1}^{s_i} [(n-j+1)+c_1(n_- - i+1)]} = \frac{q^{n_-}(1-q)^{n_+}(1+c_1)^{n_-}}{\prod_{i=1}^{n_-} \prod_{j=s_{i-1}+1}^{s_i} \left(1+c_1 \frac{n_- - i+1}{n-j+1}\right)}.$$

Thus, for $c_1 = 0$, we get correctly $P = q^{n_-}(1-q)^{n_+}$. Taking logarithms, we get

$$\log P = \log q^{n_-}(1-q)^{n_+}(1+c_1)^{n_-} - \sum_{i=1}^{n_-} \sum_{j=s_{i-1}+1}^{s_i} \log\left(1+c_1 \frac{n_- - i+1}{n-j+1}\right).$$

Now, $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$. For x very small, we may disregard all terms beyond the first one. Thus, for c_1 small enough,

$$\log P \sim \log q^{n_-}(1-q)^{n_+}(1+c_1)^{n_-} - c_1 \sum\sum \frac{n_- - i+1}{n-j+1}.$$

The most powerful rank test for the case when $\omega_2(q)$ contains only this distribution is obtained by letting ψ equal a strictly increasing function of P, say $\frac{1}{c_1}\log P$. From the above expression it is seen that for c_1, q, and n fixed, the first term is a function of n_+ only, and is an increasing function of $\psi_1 = n_+ - n_-$. Making c_1 small enough, and taking

$$\psi_2 = -\sum\sum \frac{n_- - i+1}{n-j+1}, \text{ or } = \left(\sum\sum \frac{n_- - i+1}{n-j+1}\right)^{-1} \tag{12}$$

we get a rank-admissible modified sign test family that is uniformly more powerful for all α than the sign test family, and superior in the minimax sense to any family that is not a modified sign test family.

Another modified sign test family satisfying (10) is the Wilcoxon-sign test family. It may be intuitively suggested that it is also rank-admissible, but the proof is still lacking. Computationally, it is more convenient than (12). As none of these tests have been discussed in the literature, Tables 1 and 2 give the values of c and p

Table 1. The modified sign test family generated by $\psi_2 = \left(\sum_{i=1}^{n_-} \sum_{j=s_{i-1}+1}^{s_i} \frac{n_- - i + 1}{n - j + 1} \right)^{-1}$.

n	$\alpha = 0.05$				$\alpha = 0.01$			
	$\varphi(x) = 1$ if		$\varphi(x) = p$		$\varphi(x) = 1$ if		$\varphi(x) = p$	
	either	or	if	where $p =$	either	or	if	where $p =$
5	$n_- = 0$	–	$n_- = 1$, $\psi_2 = 5$	0.60				
6	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 5$	$n_- = 1$, $\psi_2 = 4$	0.20				
7	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 3$	$n_- = 1$, $\psi_2 = 2$	0.40	$n_- = 0$	–	$n_- = 1$, $\psi_2 = 7$	0.28
8	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq \frac{840}{378}$	$n_- = 2$, $\psi_2 = \frac{840}{380}$	0.80	$n_- = 0$	$n_- = 1$, $\psi_2 = 8$	$n_- = 1$, $\psi_2 = 7$	0.56
9	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq \frac{2520}{1400}$	$n_- = 2$, $\psi_2 = \frac{2520}{1470}$	0.80	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 6$	$n_- = 1$, $\psi_2 = 5$	0.12
10	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq \frac{2520}{3240}$	$n_- = 2$, $\psi_2 = \frac{2520}{3360}$	0.20	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 2$	$n_- = 1$, $\psi_2 = 1$	0.24
11	$n_- \leq 2$	$n_- = 3$, $\psi_2 \geq \frac{27720}{22410}$	$n_- = 3$, $\psi_2 = \frac{27720}{22704}$	0.40	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq \frac{27720}{9625}$	$n_- = 2$, $\psi_2 = \frac{27720}{9660}$	0.48
12	$n_- \leq 2$	$n_- = 3$, $\psi_2 \geq \frac{27720}{30660}$	$n_- = 3$, $\psi_2 = \frac{27720}{30723}$	0.80	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq \frac{27720}{11970}$	$n_- = 2$, $\psi_2 = \frac{27720}{12474}$	0.98

for $\alpha = 0.05$ and 0.01, and for $n = 5, 6, \ldots, 12$ (test family (12)), and $n = 5, 6, \ldots, 15$ (Wilcoxon-sign test family) respectively. For large values of n, the Wilcoxon ψ_2 was shown by MANN and WHITNEY (1947) to be approximately normally distributed within each T_k with mean $\frac{1}{2}(n_+ - n_-)(n+1)$ and variance $\frac{1}{3} n_+ n_- (n+1)$. If the normal approximation is used for $11 \leq x \leq 15$, the actual size of the test will be:

$n =$	11	12	13	14	15
$\alpha = 0.05$	0.053	0.050	0.049	0.051	0.049
$\alpha = 0.01$	0.0103	0.0105	0.0098	0.0106	0.0103.

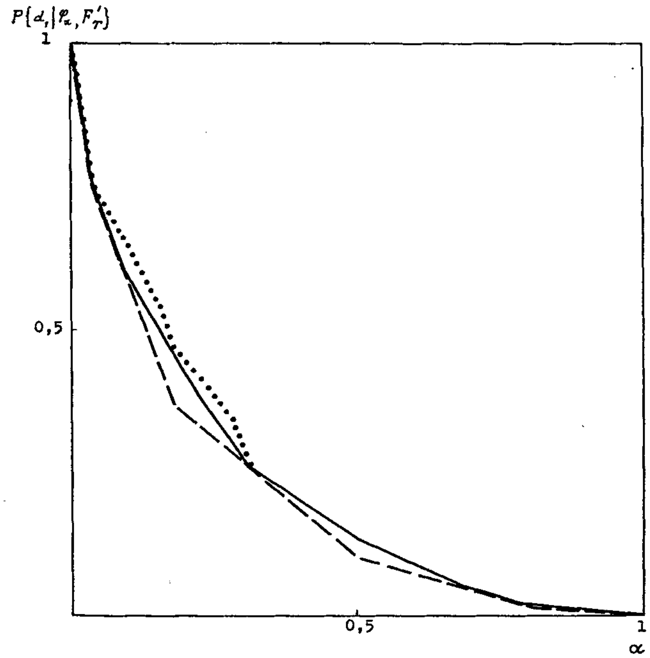
Thus, the approximation is very close already for these sample sizes, and can profitably be used.

Table 2. The Wilcoxon-sign test family, generated by $\psi_2 = \sum_{i=1}^{n} \varepsilon_i\, i$.

n	$\alpha = 0.05$				$\alpha = 0.01$			
	$\varphi(x) = 1$ if		$\varphi(x) = p$		$\varphi(x) = 1$ if		$\varphi(x) = p$	
	either	or	if	where $p =$	either	or	if	where $p =$
5	$n_- = 0$	–	$n_- = 1$, $\psi_2 = 13$	0.60				
6	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 17$	$n_- = 1$, $\psi_2 = 15$	0.20				
7	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 18$	$n_- = 1$, $\psi_2 = 16$	0.40	$n_- = 0$	–	$n_- = 1$, $\psi_2 = 26$	0.28
8	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq 28$	$n_- = 2$, $\psi_2 = 26$	0.90	$n_- = 0$	$n_- = 1$, $\psi_2 = 34$	$n_- = 1$, $\psi_2 = 32$	0.56
9	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq 29$	$n_- = 2$, $\psi_2 = 27$	0.90	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 37$	$n_- = 1$, $\psi_2 = 35$	0.12
10	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq 25$	$n_- = 2$, $\psi_2 = 23$	0.60	$n_- = 0$	$n_- = 1$, $\psi_2 \geq 37$	$n_- = 1$, $\psi_2 = 35$	0.24
11	$n_- \leq 2$	$n_- = 3$, $\psi_2 \geq 40$	$n_- = 3$, $\psi_2 = 38$	0.44	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq 54$	$n_- = 2$, $\psi_2 = 52$	0.83
12	$n_- \leq 2$	$n_- = 3$, $\psi_2 \geq 38$	$n_- = 3$, $\psi_2 = 36$	0.05	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq 56$	$n_- = 2$, $\psi_2 = 54$	0.59
13	$n_- \leq 3$	$n_- = 4$, $\psi_2 \geq 59$	$n_- = 4$, $\psi_2 = 57$	0.42	$n_- \leq 1$	$n_- = 2$, $\psi_2 \geq 53$	$n_- = 2$, $\psi_2 = 51$	0.64
14	$n_- \leq 3$	$n_- = 4$, $\psi_2 \geq 53$	$n_- = 4$, $\psi_2 = 51$	0.64	$n_- \leq 2$	$n_- = 3$, $\psi_2 \geq 75$	$n_- = 3$, $\psi_2 = 73$	0.35
15	$n_- \leq 3$	$n_- = 4$, $\psi_2 \geq 46$	$n_- = 4$, $\psi_2 = 44$	0.59	$n_- \leq 2$	$n_- = 3$, $\psi_2 \geq 76$	$n_- = 3$, $\psi_2 = 74$	0.65

As an example of the use of the tables and of the approximation, take the sequence with $n = 12$, $n_+ = 9$, and $n_- = 3$: $+ - - + + - + + + + + +$. For the Wilcoxon-sign test family $\psi_2 = 1 - 2 - 3 + 4 + 5 - 6 + 7 + 8 + 9 + 10 + 11 + 12 = 56$, and for test family (12) $\psi_2 = \left(\frac{1}{7} + \frac{2}{10} + \frac{3}{11}\right)^{-1} = \frac{27720}{16064}$. The tables show that according to both test families, the sequence is significant at the 5 % level but not at the 1 % level.

Let us also compute the normal approximation of Wilcoxon's test for $\alpha = 0.05$. From binomial tables or by easy calculation, we find that the basic cells with



——— The Wilcoxon test family
· · · · · · The Wilcoxon-Hemelrijk test family
— — — — The Wilcoxon-sign test family (or any other modified sign test family, or the sign test family).

Diagram 1. (1–power) against the bi-rectangular distribution as a function of the size, for test families using the Wilcoxon ψ_2.

$n_- = 0$, 1, and 2 give a combined probability of $\dfrac{79}{4096} = 0.0193$. The missing 0.0307 have to be taken from cells with $n_- = 3$, the total probability of which is $\dfrac{220}{4096} = 0.0537$. Thus the conditional probability within T_6 (where $n_- = 3$) should be $\dfrac{0.0307}{0.0537} = 0.572$, which in the normal distribution corresponds to a deviation from the mean of 0.18 standard deviations. Since

$$E(\psi_2) = \tfrac{1}{2}\cdot 6\cdot 13 = 39, \text{ and } D^2(\psi) = \tfrac{1}{3}\cdot 9\cdot 3\cdot 13 = 117,\ D(\psi_2) = 10.8,$$

we get the approximate value of c to be $39 - 0.18\cdot 10.8 = 37.1$. Since the observed value of $\psi_2 = 56$ is greater than 37.1, we get $\varphi(x) = 1$, i.e. the hypothesis of zero mean is rejected on the 5 % level.

As a corollary to theorem 8, all test families with ψ_2 satisfying (10) have the bi-rectangular for least favorable distribution. This is true of the Wilcoxon tests 2a, b, and c. By Theorem 7, the Wilcoxon-sign test family is superior to the others, and since the pure Wilcoxon test is closer to the ψ_1-order, it is superior to the Wilcoxon-Hemelrijk test family. As an example, Diagram 1 gives $P(d_1|\varphi_\alpha, F'_T)$ as a function of α for $n = 5$, $q = \frac{1}{4}$, and $F'_T =$ the bi-rectangular distribution.

In order to give some illustration to the question of what the restriction to rank tests means, it is of interest to compare the pure Wilcoxon test family with the permutation test family generated by $\psi(x)=\bar{x}$("Fisher's tests"). These were investigated by LEHMANN and STEIN (1949), and as a special case of their results, Fisher's tests are uniformly most powerful among permutation tests (including rank tests) when Ω' contains only normal distributions. According to HOEFFDING (1952), they are also asymptotically as powerful as the t tests in the same class. However, in Ω'_E, neither Wilcoxon's tests nor Fisher's tests are for all α uniformly more powerful than the others. For take any distribution that has the bi-rectangular for its transform. Then the scheme of p. 148 may be extended in a slightly modified form to Fisher's test family, still considering only the case $n=5$.

Take first $\alpha=5/32$. From p. 148 it is seen that Wilcoxon's test (2a) will have $\varphi(x)=1$ for B_{51}, B_{31}, B_{32}, B_{33}, and B_{11}. In Fisher's test, B_{51}, B_{31}, B_{32}, and B_{33} will always have $\varphi(x)=1$, irrespective of the absolute values of the observations. The fifth basic cell with $\varphi(x)=1$ will be B_{11} if $y_1+y_2<y_4$, otherwise B_{34} will take its place. Thus, since $P(B_{34}|F)>P(B_{11}|F)$ for all F with the bi-rectangular distribution as transform, Fisher's test is more powerful for these distributions than Wilcoxon's test. On the other hand, take $\alpha=10/32$. Both tests have $\varphi(x)=1$ for B_{51}, B_{31}, B_{32}, B_{33}, B_{34}, B_{11}, and B_{12}. The following scheme will show which other cells enter the region where $\varphi(x)=1$.

	Wilcoxon's test	Fisher's test
If $\max[(y_1+y_4),\, y_5]<\min[(y_1+y_2+y_3),\, (y_2+y_4)]$	B_{35}	B_{35}
$\max[(y_1+y_2+y_3),\, y_5]<y_1+y_4$	B_{35}	B_{35}
$y_2+y_4<\min[(y_1+y_2+y_3),\, y_5]$	B_{35}	B_{16}
$\max[(y_1+y_4),\, (y_1+y_2+y_3)]<\min[(y_1+y_4),\, y_5]$	B_{35}	B_{11}
$\max[(y_1+y_4),\, y_5]<y_1+y_2+y_3$	B_{13}	B_{13}
$\max[(y_1+y_4),\, (y_1+y_2+y_3)]<y_5$	B_{13}	B_{13}
$\max[(y_1+y_2+y_3),\, y_5]<y_1+y_4$	B_{13}	B_{-11}
$y_2+y_3<y_1+y_5$	B_{14}	B_{14}
$y_1+y_5<y_2+y_3$	B_{14}	B_{15}

The events within each group are complementary. It is seen that for some sample points (which have positive probability) Fisher's test takes basic cells from a T_k with lower k, and thus with smaller probability according to the bi-rectangular distribution, than does Wilcoxon's test. As for any q-ordered incomplete weight function the bi-rectangular distribution is least favorable relative to the Wilcoxon test family, Fisher's test of size $\alpha=10/32$ cannot be conditionally better in the minimax sense than Wilcoxon's. As a consequence, Fisher's test family is not superior to Wilcoxon's test family with respect to such weight functions, and a fortiori not superior to the admissible modified sign test families.

Summing up the results of this section, the sign test family is, according to the criterion adopted here, better than any other test taken into consideration, if it is not known that $P(-\,+)\geq P(+\,-)$. On the other hand, if the distribution is known to have a density function that is logarithmically concave downwards, a modified sign test family satisfying (10) is best among rank test families. This strong position of the modified sign test families is rather unexpected, as they have not, to the author's knowledge, been investigated before. Their optimality property is destroyed if Ω' is further reduced in such a way that the bi-rectangular distribution is left out. Only in very rare cases, however, does the information permit such a reduction, except when the functional form of the distribution is known, and in that case ordinary parametric tests should be used.

REFERENCES

BLOMQVIST, N. (1950), On a measure of dependence between two random variables. AMS *21*, p. 593–600.

COCHRAN, W. G. (1937), The efficiencies of the binomial series tests of significance of a mean and of a correlation coefficient. JRSS *100*, p. 69–73.

DIXON, W. J. and MOOD, A. M. (1946), The statistical sign test. JASA *41*, p. 557–566.

FISHER, R. A. (1925), Statistical methods for research workers, Sec. 24, ex. 19. Edinburgh.

—— (1935), Design of experiments, Sec. 21. Edinburgh.

HEMELRIJK, J. (1950a), A family of parameterfree tests for symmetry with respect to a given point. Proc. Kon. Ned. Akad. van Wetensch. *53*, p. 945–955, 1186–1198.

—— (1950b), Symmetritoetsen en andere toepassingen van de theorie van Neyman en Pearson. Amsterdam.

HOEFFDING, W. (1951), "Optimum" nonparametric tests. 2nd Berkeley Symp. p. 83–92.

—— (1952), The large-sample power of tests based on permutations of observations. AMS *23*, p. 169–192.

LEHMANN, E. L. (1947), On families of admissible tests. AMS *18*, p. 97–104.

—— (1953), The power of rank tests. AMS *24*, p. 23–43.

LEHMANN, E. L. and STEIN, C. (1949), On the theory of some non-parametric hypotheses. AMS *20*, p. 28–45.

LINDLEY, D. V. (1953), Statistical inference. JRSS B *15*, p. 30–76.

MANN, H. B. and WHITNEY, D. R. (1947), On a test of whether one of two random variables is stochastically larger than the other. AMS *18*, p. 50–60.

NEYMAN, J. and PEARSON, E. S. (1936), Contributions to the theory of testing statistical hypotheses, Part I. Statistical Research Memoirs, Vol. I, London.

PITMAN, E. J. G. (1948), Notes on non-parametric statistical inference. Mimeogr., Columbia Univ.

SVERDRUP, E. (1952), Weight functions and minimax procedures in the theory of statistical inference. Arkiv for Mathematik og Naturvidenskap, *LI* No. 7. Oslo.

—— (1953), Similarity, unbiassedness, minimaxibility and admissibility of statistical tests. SA *36*, p. 64–86.

VAN DER VAART, H. R. (1950), Some remarks on the power function of Wilcoxon's test for the problem of two samples. Proc. Kon. Ned. Akad. van Wetensch. *53*, p. 494–520.

WALD, A. (1942), On the principles of statistical inference. Notre Dame Mathematical Lectures, No. 1.

—— (1950), Statistical decision functions, New York.

WALSH, J. E. (1946), On the power function of the sign test for slippage of means. AMS *27*, p. 358–362.

—— (1949), Some significance tests for the median which are valid under very general conditions. AMS *20*, p. 64–81.

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

WOLFOWITZ, J. (1942), Additive partition functions and a class of statistical hypotheses. AMS *13*, p. 247–279.

Tryckt den 26 april 1954

Uppsala 1954. Almqvist & Wiksells Boktryckeri AB