A Comparison of Tests for Ordered Alternatives

With Application in Medicine

By

Lisbeth Alainentalo

Bachelor's Thesis
Department of Mathematical Statistics
Umeå University
1997

Supervisor: Staffan Uvell
Summary

A situation frequently encountered in medical studies is the comparison of several treatments with a control. The problem is to determine whether or not a test drug has a desirable medical effect and/or to identify the minimum effective dose. In this Bachelor’s thesis, some of the methods used for testing hypotheses of ordered alternatives are reviewed and compared with respect to the power of the tests. Examples of multiple comparison procedures, maximum likelihood procedures, rank tests and different types of contrasts are presented and the properties of the methods are explored.

Depending on the degree of knowledge about the dose-responses, the aim of the study, whether the test is parametric or non-parametric and distribution-free or not, different recommendations are given which of the tests should be used. Thus, there is no single test which can be applied in all experimental situations for testing all different alternative hypotheses.
Introduction

A common problem in medical studies is to determine whether or not a test drug has a desirable biological effect. One objective of such a study is to test if there exist an ordered treatment effect among different doses of a test drug and/or to identify the lowest dose of the drug that will cause a desirable effect. In this type of experiments several doses in increasing order of the medicine are distributed to separate groups and compared with a control or placebo group. The treatment units are randomly allocated to one of $k$ groups representing each dose level or to a control group. The mean response of the treatments is denoted by $\mu_i$ where $\mu_0$ is the control group mean and $\mu_i$ ($i = 1, 2, \ldots, k$) is the $k$ dose-group means.

In this paper, the null hypothesis is always

$$H_0 : \mu_0 = \mu_1 = \ldots = \mu_k$$

i.e. there is no treatment effect.

Depending on the object of the medical study and the degree of knowledge about the effect of the test drug, the alternative hypothesis may be formulated in different ways.

If nothing is known about the shape of the response or if the $k$ groups representing different treatments cannot be ordered, the alternative hypothesis is

$$H_1 : \mu_0 \leq \mu_i \quad i = 1, 2, \ldots, k$$

with at least one strict inequality. Bartholomew (1961) called this the simple tree order alternative hypothesis.

On the other hand, if we are interested in whether the responses increase monotonically with dose level, the alternative hypothesis is

$$H_1 : \mu_0 \leq \mu_1 \leq \ldots \leq \mu_k$$

with at least one strict inequality. This hypothesis is often referred as simple order alternative.

In many cases a more specific alternative hypothesis than (1) is more reasonable to the researcher but still not as strict as (2). A set of hypothesis that lie between these two extremes is

$$H_1 : \mu_0 = \mu_1 = \ldots = \mu_i < \mu_j \quad j = i+1, \ldots, k \text{ for any given } i < k$$

This hypothesis is used when the researcher is interested in finding the minimum effective dose (MED) (Ruberg 1989).

However, if it is known that the treatment effects are monotonically increasing up to a point followed by a monotonic decrease due to toxic effects of the drug at high doses, the treatment effects are said to follow an umbrella pattern (Mack & Wolfe 1981). The alternative hypothesis in this case is

$$H_1 : \mu_1 \leq \ldots \leq \mu_p \geq \ldots \geq \mu_k \text{ for some } p, 1 \leq p \leq k,$$

with at least one strict inequality.
Since the early 1950s several methods have been developed to test the null hypothesis of no treatment effect against above described ordered alternative hypotheses. The requirement of controlling the type 1 error rate, i.e probability of rejecting a true $H_0$, has been one of the reasons to continue the development of the different tests. Tukey (1985) introduced the commonly used terminology trend test which is defined as "test for progressiveness of response with increasing dose including controls as zero dose". To avoid misunderstandings the term test for ordered alternatives will be used consequently throughout this paper since it also include the term trend test.

The purpose of this paper is to give a chronological summerize and comparison of the methods used for testing hypotheses of ordered alternatives; multiple comparison procedures, maximum-likelihood procedures, rank test and contrast procedures. The properties of the methods are explored and compared with respect to the power of the tests.

**Literature suvey**

All information in this paper has been obtained by help of Current Index to Statistics, 1995 edition. The main search terms were: ordered alternatives, trend test, increasing dose, multiple comparison and dose finding.

**Tests for ordered alternatives**

Several different methods and procedures have been developed to test for ordered alternatives. An ordered alternative is a hypothesis that specifies a particular ordering of the $\mu_i$ prior to observation of the data where $\mu_i$ ($i = 0, 1, \ldots, k$) represent location parameters e.g. treatment means for $k$ populations. In this paper, a review of some of the methods used will be presented. Firstly, a multiple comparison procedure, namely Dunnett’s procedure, is described, followed by the most commonly used maximum likelihood procedures namely; Bartholomew’s (1959, 1961) test and Williams’ (1971, 1972) test. Finally, some examples of rank tests and different types of contrasts are discussed and some examples of step up and step down procedures are briefly reviewed.

**Multiple comparison procedure**

A situation frequently encountered in medical/drug trials is the comparison of several treatments with a control or standard. The problem is to decide whether the treatments are better than the control or not. A large number of procedures have been proposed for problems involving several treatments for instance Duncan’s, Newman-Keul’s and Tukey’s test but none of these procedures take consideration of a control group. Dunnett’s (1955) multiple comparison procedure is one of the best known tests that include a control group and will be summerized below.

**Dunnett’s multiple comparison procedure**

Dunnett (1955) proposed a test that rejects $H_0$ for large values of

$$D_i = \max_{1 \leq i \leq k} \frac{X_i - \bar{X}_0}{S \sqrt{\frac{1}{n_i} + \frac{1}{n_0}}}$$
where

\[ \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad S^2 = \frac{1}{k} \sum_{i=0}^{k} \left( \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \right) / (N - k - 1), \quad N = \sum_{i=0}^{k} n_i \]

and \( X_{ij} \) (\( j = 1, \ldots, n_i; \ i = 0, \ldots, k \)) are independent normal observations with means \( \mu_i \) and common variance \( \sigma^2 \). The critical values for equal sample sizes and one-sided comparisons were tabulated by Dunnett in 1955 and tables for two-sided comparisons were published in 1964.

Dunnett’s test is appropriate to test \( H_0 \) versus \( H_1 \) - \( H_0 \) i.e. \( H_1 \) but not \( H_0 \), where \( H_1 \) is the simple tree alternative (1), but it can also be used for identifying the MED (Ruberg 1989). If \( m \) is the smallest index for which \( D_i \) is significant then the \( m \)th dose level will be MED. When all the treatment means are equal, Dunnett’s test has good power (Murkerjee, 1987) but its power function vary considerably over alternative regions. If one has no further information indicating a proper alternative hypothesis other test would be preferred.

**Maximum likelihood procedures**

Now, two of the most commonly used maximum likelihood procedures namely; Bartholomew’s and Williams’ test, will be discussed. These tests estimate the treatment means under the alternative hypothesis (2) by the method of maximum likelihood.

**Bartholomew’s test for ordered means**

Bartholomew (1959) constructed a likelihood ratio test statistic for testing the null hypothesis against the alternative hypothesis \( H_1 : \mu_0 \geq \mu_1 \geq \ldots \geq \mu_k \) (decreasing order) with at least one strict inequality for which the rank order of the means is known. He proposed the statistic

\[ X^2_k = \sum_{i=1}^{k} (\mu^*_i - \bar{X})^2 \quad \text{where} \quad \bar{X} = \frac{1}{k} \sum_{i=1}^{k} x_i / k \]

and the independent observations \( X_i \) (\( i = 1, \ldots, k \)) are assumed to be normally distributed with mean \( \mu_i \) and with known standard deviation \( \sigma^2 \), and \( \mu^*_i \) is the value of \( \mu_i \) which maximize (5) subject to the condition \( \mu_1^* \geq \mu_2^* \geq \ldots \geq \mu_k^* \).

In 1961, Bartholomew extended his test statistic to include the case of unknown standard deviation \( \sigma^2 \) and provided extended tables of significant points for equal and unequal group sizes. The test statistic is

\[ E^2_k = \sum_{i=1}^{k} n_i (\mu^*_i - \bar{x})^2 / \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \quad \text{where} \quad \bar{x} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} / \sum_{i=1}^{k} n_i \]

\( \mu^*_i \) is the maximum likelihood estimate of the \( i \)th group mean, \( k \) is the number of groups and \( n_i \) the size of \( i \)th group.

This test can be applied in a stepwise manner, firstly applying it to all the treatment groups then, if a treatment effect is found, exclude this dose level mean and applying the test to the rest of the means and so on.
Williams’ test

Williams (1971, 1972) considered the case in which one assume not only that the treatment means are greater than the control but also that the ordering among the treatment means is completely known. One application of this order restriction on the $k$ treatment means and the control mean is the comparison of increasing dosage levels of a drug with a zero-dose control. Williams’ test assumes that the dose effect, if any, satisfies (2) and can be applied in a stepwise fashion to determine the MED.

The maximum likelihood estimates of the treatment means $\mu_i$ ($i = 1, \ldots, k$) are obtained analytically from the sample means, $\bar{x}_i$, and the number of observations in the samples, $n_i$, by the formula

$$\mu_i^* = \max_{1 \leq u \leq i} \min_{i \leq v \leq k} \left( \sum_{j=u}^{n_j} \bar{x}_j / \sum_{j=u}^{n_j} n_j \right) \quad i = 1, \ldots, k$$

The statistic

$$t_k = \frac{\mu_k^* - \bar{x}_0}{(s^2 / n_k + s^2 / n_0)^{1/2}}$$

can be calculated, where $s^2$ is the estimate of the residual variance, $n_0$ is the number of observations in the control group and $\bar{x}_0$ is the control group sample mean. Williams’ test is designed for the analysis of normally distributed data with equal group variances. The statistic $t_k$ is then compared with the value of $t_k^*$ obtained from the tables provided by William (1972). If the the statistic $t_k$ is significant then the statistic $t_{k-1}$ is calculated in the same way and compared with the value of $t_{k-1}^*$ given in the tables. This procedure continues until a non-significant $t_{n-1}$ is obtained for some dosage level, $m-1$. Since $\mu_m^*$ is the smallest dose group mean that is declared different from the control mean response the $m$th dose group is the MED. In other words, the estimate of MED, $\text{MED}^* = m$, if $m$ is the smallest index for which $H_0$ is rejected. If there is no such $m$ then conclude that there is no dose level as the MED or MED is higher than the dose level $k$.

Williams compared the power of his test with Bartholomew’s test (1961) and Dunnett’s test for comparing several treatments with a control (1955, 1964). He found his test to be superior to Dunnett’s test, but slightly less powerful than Bartholomew’s test which becomes more powerful as $k$ increases. Bartholomew’s test is generally more powerful than Williams’ test when four or more groups are compared. This however refers only to the detection of effects at the highest dose level. Williams’ test is generally the superior when only two or three groups are compared.

A tests based on modifications to Williams’ test is proposed by Shirley (1979) and is discussed later on page 11 in this paper.

Rank tests

One of the simplest questions that can be asked in a comparison of $k$ treatments is whether there is any difference among these treatments. If the difference between treatments is sufficiently large, it will be reflected in the rank averages.

Let $n_i$ be the sample size and $\bar{R}_i$ be the average rank for the $i$th group, then the most commonly used procedure, the Kruskal-Wallis (1952), is based on the statistic
\[ W = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i (\bar{R}_i - (N + 1)/2)^2 \quad \text{where} \quad N = \sum_{i=1}^{k} n_i \]

Now, suppose that the treatments are ordered, of course before the responses have been observed. Then, the Kruskal-Wallis test is no longer appropriate since it rejects \( H_0 \) whenever the rank averages \( \bar{R}_i \) are sufficiently different, regardless of their order. Testing equality against ordered alternatives using ranks has been the subject of many papers: among them Terpstra (1952), Jonkheere (1954), Chacko (1963) and Le (1988). The tests proposed by these authors will be discussed below.

**Jongheere’s (or Jongheere-Terpstra’s) test for ordered alternatives**

Two of the earliest works in the area of testing for ordered alternatives were made by Terpstra (1952) and Jonkheere (1954). Here it is assumed that the ordering is \( \mu_1 < \mu_2 < \ldots < \mu_k \). Jongheere recognized the similarity to problems of monotone trend and developed tests for ordered alternatives in the one-way layout based on Kendall’s test for rank correlation. Terpstra presented a test that is equivalent to Jongheere but in a slightly different form. This test is in the literature known as Jongheere’s test or as Jongheere-Terpstra’s test.

In the original article by Jongheere a statistic \( S \) was proposed to test the hypothesis that the \( k \) populations are identical against the alternative that the populations are stochastically ordered in a specified manner i.e.

\[
H_0: F_1(x) = F_2(x) = \ldots = F_k(x) \quad \text{for all} \ x \ \text{versus} \\
H_1: F_1(x) \geq F_2(x) \geq \ldots \geq F_k(x) \quad \text{with at least one strict inequality for some} \ x \ \text{where} \ F_i \ \text{is cumulative distribution functions (cdf)}
\]

**Test procedure**

Assume we are given random samples of size \( n_1, n_1, n_2, \ldots, n_k \) respectively from each of the \( k \) populations. Denote by \( X_{ij} \) the \( j \)th observation in the sample from the \( i \)th population \( (j = 1, \ldots, n_i, i = 1, \ldots, k) \). Denote \( F_i \) the continuous cdf of \( X_{ij} \). For \( i > 1 \), define \( \phi(X_{ij}) \) to be the number of observations from the first \( (i-1) \) populations which are less than \( X_{ij} \). Let

\[
S_i = \sum_{j=1}^{n_i} \phi(X_{ij}) \quad \text{and} \quad S = \sum_{i=2}^{k} S_i
\]

Test the null hypothesis, \( H_0 \) versus the alternative hypothesis, \( H_1 \). \( H_0 \) is rejected if \( S \geq S_\alpha \) where \( S_\alpha \) is tabulated critical value. Jonkheere proposed the test statistic of the form \( 2S - M \) where \( M \) is the maximum possible value of \( S \).

The computation of Jongheere’s statistic is simple and straightforward. The data is arrayed in \( k \) treatment columns in the left-to-right order implied by the alternative hypothesis. Then
where \( u \) and \( w \) are treatment subscripts with ranges \( u = 1, \ldots, k-1 \) and \( w = u+1, \ldots, k \) and \( t \) and \( v \) are observation subscripts with ranges \( t = 1, \ldots, n_u \) and \( v = 1, \ldots, n_w \). The sum of the \( \phi(X_{tu}, X_{vw}) \) over the ranges of all the subscripts is the test statistic with larger values favouring rejection of the null hypothesis.

Although Jonkheere developed this distribution-free and non-parametric test as an extension of Kendall’s test for rank correlation a more convenient equivalent formulation is now used. If we assume \( F_i(x) = F(x - \mu_i) \) for some unknown cdf, \( F \), the problem becomes

\[
H_0: \mu_1 = \mu_2 = \ldots = \mu_k \quad \text{versus} \quad H_1: \mu_0 < \mu_1 < \ldots < \mu_k
\]

The modified Jonkheere statistic \( J \) (Potter and Sturm, 1981) is defined as

\[
J = \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} U_{uv}
\]

where \( U_{uv} \) is the Mann-Whitney count between samples \( u \) and \( v \) i.e.

\[
U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi(X_{iu}, X_{jv}) \quad \text{where} \quad \phi(X_{iu}, X_{jv}) = \begin{cases} 1 & \text{if } X_{iu} < X_{jv} \\ 0 & \text{otherwise} \end{cases}
\]

For large values of \( J \) the null hypothesis is rejected in favor of the alternative hypothesis. Tabulated values are published by Jonkheere (1954) and Odeh (1971).

In the case of large samples, when \( F \) is not continuous or when ties exist (e.g. due to round off) an approximate test is used based on the fact that \( J \) is asymptotically normally distributed when \( \min_i \{n_i\} \to \infty \) (Hollander and Wolfe, 1973, tables and formulas). The approximate test rejects \( H_0 \) in favor of the à priori ordering when the right-tail \( \alpha \)-level critical value of the standard normal is exceeded by

\[
Z_J = \frac{J - \mu_J}{\sigma_J}
\]

The exact null moments of \( J \) are given by

\[
\mu_J = \frac{(N^2 - \sum_{i=1}^{k} n_i^2)}{4}
\]

\[
\sigma_J^2 = \frac{[N^2(2N + 3) - \sum_{i=1}^{k} n_i^2(2n_i + 3)]}{72} \quad \text{and} \quad N = \sum_{i=1}^{k} n_i
Nelson and Toothaker (1975) investigated the power of Jonkheere’s test and found that the test is sensitive to ordered location differences and it is also sensitive to general cumulative distribution differences. They recommended to use the Jonkheere’s test to test hypothesis concerning ordered distributions.
Other tests and comparisons

Among other rank tests developed for ordered alternatives in the one-way layout two are described below; Chacko’s test and Le’s test, both using the treatment rank sums \( R_1, \ldots, R_k \) based on the combined ranking of all N observations.

Chacko’s test

Chacko (1963) proposed a rank test similar to the test by Kruskal-Wallis for stochastic ordering of populations for the case of equal sample sizes. Let \( k \) independent random samples of equal size \( n \) be drawn from \( k \) univariate population with unknown cumulative distributions \( F_i (i = 1, \ldots, k) \) respectively. It is assumed that each \( F_i \) is continuous to avoid problem of ties. The hypothesis that the \( k \) populations are identical is tested against the alternative that the populations are stochastically ordered.

Chacko (1963) proposed a test procedure for ordered alternatives:

Let \( R_1^* \leq \ldots \leq R_k^* \) be the isotonic regression of the average ranks \( \bar{R}_1, \ldots, \bar{R}_k \) under the order restriction \( \mu_1 \leq \ldots \leq \mu_k \). For a discussion of the algorithm for obtaining \( R_1^*, \ldots, R_k^* \) see Barlow (1972). Chacko’s rank test rejects \( H_0 \) for large values of

\[
\bar{X}^2_k = \frac{12}{N+1} \sum_{i=1}^{k} \lambda_i \left( R_i^* - \frac{(N+1)}{2} \right)^2
\]

where \( \lambda_i = n_i / N \) (i = 1, ..., k).

When \( k=2 \) the test is the same as the one-tail test of Wilcoxon. The test based on Chacko’s statistics, \( X^2_k \), show larger power than Jonkheere’s test when there is considerable variation in the differences between consecutive means but Chacko’s test is only valid for equal sample sizes.

Le’s test

The distribution free test against ordered alternatives proposed by Jonkheere is based on the Kendall’s rank correlation. Le (1988) proposed a test based on Spearman’s rank correlation instead of Kendall’s and the test has similar functional structure as the Kruskal-Wallis test.

Le proposed the statistic

\[
W = \sum_{i=1}^{k} n_i (L_i - M_i) \bar{R}_i, \quad \text{where} \quad L_i = \sum_{t<i} n_t, \quad M_i = \sum_{t>i} n_t
\]

and \( n_i \) is sample size of treatment group \( i \) (i = 1, ..., k) and \( \bar{R}_i \) the average rank for the \( i \)th group (j = 1, ..., \( n_i \)). The statistic \( W \) is a linear contrast among rank averages \( \bar{R}_i \), i.e.

\[
\sum_{i=1}^{k} n_i (L_i - M_i) = \sum_{i=1}^{k} \lambda_i = 0 \quad \text{with coefficients} \quad \lambda_i = n_i (L_i - M_i)
\]
The exact null moments of $W$ are

$$\mu_w = \frac{N + 1}{2} \sum_{i=1}^{k} \lambda_i = 0 \quad \sigma^2_w = \frac{N(N + 1)}{12} \sum_{i=1}^{k} n_i (L_i - M_i)^2 \quad (6)$$

If $\min\{n_i\} \to \infty$ the statistic is approximately normally distributed with null mean zero and null variance given in (6). The test based on $W$ possesses two important optimality properties. When $\mu_i$ are equally spaced and $F$ is logistic cdf, Hajek and Sidák (1967) showed $W$ to be locally most powerful rank test and to be asymptotically most powerful (fixed $k$, $\min\{n_i\} \to \infty$) among all rank tests.

**Shirley’s test**

Shirley (1977) suggested a non-parametric version of Williams’ (1971, 1972) test. As in Williams’ test, one is interested in comparing increasing doses of a test substance with a control and also, one uses the prior information that the responses to the substance are monotonically ordered. Shirley’s test can be used to determine the lowest dose level at which there is evidence of a difference from the control.

**Test procedure**

Let $n_0$ be sample size of control group and $n_i$ be sample size of treatment group $i$ ($i = 1, \ldots, k$). Calculate the average rank for the $i$th treatment group $\bar{R}_i$ where $\bar{R}_0$ is the average rank for the control group. Shirley proposed the test statistic for the case of equal sample sizes

$$t_k = \frac{\max_{1 \leq u \leq k} \sum_{j=u}^{k} \bar{R}_j / (k - u + 1) - \bar{R}_0}{[(k+1)(N+1)/6]^{1/2}}$$

where $N = \sum_{i=0}^{k} n_i$

which is approximately distributed as Williams’ test statistic, $t_k^*$, with infinite degrees of freedom. Tables for $t_k^*$ given by Williams’ (1972).

Thus, the statistic $t_k$ is compared with the value of $t_k^*$ and if $t_k$ is significant then the statistic $t_{k-1}$ is calculated and compared with the value of $t_{k-1}^*$. This procedure continues until a non-significant $t_i$ is obtained for some treatment group $i$ that is, the dose level $i$ is concluded as the MED.

If the sample size of the control group is different from the sample sizes of the treatment groups, the statistic

$$t_k = \frac{\max_{1 \leq u \leq k} \left( \sum_{j=u}^{k} n_j \bar{R}_j / \sum_{j=u}^{k} n_j \right) - \bar{R}_0}{\left( N(N+1)/12 \left( \frac{1}{n_k} + \frac{1}{n_0} \right) \right)^{1/2}}$$

should be used. This statistic is also distributed as the $t_k$ statistic with infinite degrees of freedom.
If ties occur in the rankings, calculate the average ranks and make the corrections for ties by replacing the term $N(N+1)/12$ by

$$\sum_s \frac{(t_s^2 - t_s)}{12(N - 1)}$$

where there are $s$ groups of ties with $t_s$ observations tied in the $s$th group.

Compared with Williams’ test, Shirley’s test uses the mean ranks of each group instead of sample means and the estimated variance $s^2$ is replaced by

$$\frac{N(N + 1)}{12} - \sum_s \frac{(t_s^2 - t_s)}{12(N - 1)}.$$

**Rank tests for umbrella pattern**

Non-parametric procedures for comparing several treatments with a control in a one-way layout have been studied extensively. However, many of these procedures do not use any prior information about the pattern of treatment effects.

Usually, in a drug study increasing dosage levels may be compared with a zero-dose control or placebo. It is believed that the higher the dose of the drug is applied the higher will be the treatment effect. However, often it is known that the subject may suffer from toxic effect at high doses thereby decreasing the treatment effects. In this case, an ordering in the treatment effect that is monotonically increasing up to a point followed by a monotonic decrease is foreseen. This increasing-decreasing ordering is said to follow an umbrella pattern (Mack and Wolfe, 1981). The point that separates the ordering is called the peak of the umbrella.

**Mack and Wolfe’s test**

Let $X_{ij}$, $(j = 1, \ldots, n_i; i = 1, \ldots, k)$ be $k$ independent random samples with $X_{ij}$ having absolutely continuous distribution functions $F_i(x)$. One is often interested in testing the hypothesis that all $k$ samples come from a single common distribution. If the alternative hypothesis is that at least two of the $k$ distributions have different medians, a well-known distribution-free test procedure is that of Kruskal-Wallis (1952). But if a priori information about the possible alternatives is available it would be better to choose a test that is designed for detecting these alternatives. Jonkheere (1954) and Terpstra (1952) were first to consider the case of monotonically ordered alternatives. In 1981, Mack and Wolfe designed a distribution-free $k$-sample rank test for the alternative hypothesis

$$H_1 : F_1(x) \geq \ldots \geq F_p(x) \leq \ldots \leq F_k(x)$$

with at least one strict inequality for at least one $x$.

This type of alternative is appropriate when evaluating response to increasing dosage levels where the treatment effects are increasing up to a point and then decreasing due to for example toxic effects of the drug. Mack and Wolfe (1981) proposed two different tests for the case where the peak of the umbrella is known and where the peak is unknown à priori.
Test procedure in the case of known umbrella peak

Reject $H_0$ for large values of the test statistic

$$A_p = \sum_{p=1}^{k} \sum_{j=1}^{p} U_{ij} + \sum_{p=1}^{k} \sum_{j=p+1}^{k} U_{ji}$$

(7)

where $U_{uv}$ is the Mann-Whitney statistic between $u$th and $v$th samples.

Asymptotic properties

If each $n_i \to \infty$ as $N=\sum n_i \to \infty$ (i = 1, ..., k) such that

$$(n_i / N) \to \rho_i \quad \text{with} \quad 0 < \rho_i < 1 \quad \text{for} \quad i = 1, \ldots, k \quad \text{then} \quad [A_p - \mu_0(A_p)] / \sigma_0(A_p)$$

has an asymptotic null distribution that is standard normal where

$$\mu_0(A_p) = \left[ N_1^2 + N_2^2 - \sum_{i=1}^{k} n_i^2 - n_p^2 \right] / 4 \quad \text{and}$$

$$\sigma_0^2(A_p) = \frac{1}{72} \left\{ 2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2 (2n_i + 3) - n_p (2n_p + 3) + 12n_p N_1 N_2 - 12n_p^2 N \right\}$$

with $N_1 = \sum_{i=1}^{p} n_i$ and $N_2 = \sum_{i=p+1}^{k} n_i$

The complete exact null distribution of $A_p$ is presented in the article (Mack and Wolfe, 1981) for different values values of $k$, $p$ and $n_i$.

Test procedure in the case of unknown umbrella peak

Mack and Wolfe (1981) proposed a test procedure where one first use the data to estimate the unknown umbrella peak $p$ and then use the corresponding statistic given in (7). Thus, reject $H_0$ for large values of the statistic

$$A_{p^*} = \sum_{i=1}^{k} \Gamma_i (A_i - \mu_0(A_i)) / \sigma_0(A_i)$$

where $A_i$ is the peak-known-statistic given in (7) with the peak at the $i$th group, $\mu_0(A_i)$ and $\sigma_0(A_i)$ are the corresponding null mean and variance and the random variables $\{\Gamma_1, \ldots, \Gamma_k\}$ indicate which group(s) has been estimated by the samples to be the peak group(s). Their values are determined by the following procedure. Let

$$Z_t = \sum_{i=t}^{k} U_{ti} \quad \text{for} \quad t = 1, \ldots, k.$$

Note that $Z_t$ is actually a two-sample Mann-Whitney statistic computed between the $t$th sample and the remaining $k-1$ samples.
Define
\[ Z^*_i = \frac{(Z_i - \mu_0(Z_i))}{\sigma_0(Z_i)} \] where
\[ \mu_0(Z_i) = \frac{n_i(N - n_i)}{2} \]
and
\[ \sigma^2_0(Z_i) = \frac{n_i(N - n_i)(N + 1)}{12} \]
are the null mean and variance of \( Z_i \).

Let \( r \) \( (1 \leq r \leq k) \) equal the number of populations tied for having the largest \( Z^*_q \) sample value. Set \( \Gamma_i = 1/r \), if the \( r \)th population is among those tied for the largest \( Z^*_q \) value and 0 otherwise.

A Monte Carlo study demonstrated that the \( A_{p*} \) procedure is superior to either the Kruskal-Wallis test or to the Jonkheere’s test for umbrella alternatives with unknown peak \( p \). When the peak is known à priori the \( A_p \) statistic is even better but if there is any doubts about the true peak the \( A_{p*} \) statistic should be used since the power of \( A_p \) can be low when an incorrect peak is chosen.

**Chen and Wolfe’s test 1**

One should be aware of that, to ensure the distribution-free property, the Mack and Wolfe’s (1981) test require the assumption that the continous populations have the same shape. Chen and Wolfe (1990a) proposed a rank test based on modification of the Mack-Wolfe test without making the assumption that the underlying populations have the same shape (under null hypothesis). In both these tests, the expected values are the same, but when the underlying populations have different shapes the variances are changed. Chen and Wolfe’s (1990a) found the respective variances of \( Z_i \) and \( A_p \) \( (t = 1, \ldots, k) \) (when not assuming that the populations have the same shape) to be

\[ \text{Var } Z_i = \sum_{t=1}^{k} (w_{it}^2 + w_{it}^2 + P_{it}^2 - P_{it}^2) + 2 \sum_{i<j} \sum_{t=1}^{k} w_{ijt} \quad \text{and} \]

\[ \text{Var } A_p = \sum_{t=1}^{k} \sum_{j=1}^{k-1} (w_{ij}^2 + w_{ij}^2 + P_{ij}^2 - P_{ij}^2) + \sum_{i<j} \sum_{t=1}^{k} (w_{ij}^2 + w_{ij}^2 + P_{ij}^2 + P_{ij}^2) + \]

\[ + 2 \left( \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} \sum_{t=1}^{k} (w_{ijt}^2 + w_{ijt}^2 - w_{ijt}^2) + \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} \sum_{t=1}^{k} (w_{ijt}^2 + w_{ijt}^2 - w_{ijt}^2) + \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} \sum_{t=1}^{k} w_{ijt} \right) \]

where \( t = 1, \ldots, k; \)

\[ w_{ijt} = \sum_{v=1}^{n} (P_{ij} - \bar{P}_{ij})(P_{ij} - \bar{P}_{ij}) \quad i, j, t = 1, \ldots, k \]

\[ P_{ij} = \sum_{v=1}^{n} \psi(X_{jv} - X_{iv}) \quad v = 1, \ldots, n_j \quad \text{and} \quad \bar{P}_{ij} = \sum_{v=1}^{n} P_{ij} / n_j \quad \text{where} \quad \psi(a) = \begin{cases} 1 & \text{for } a \geq 0 \\ 0 & \text{for } a < 0 \end{cases} \]

For the case of the known peak umbrella alternative, \( H_0 \) is rejected for large values of
\[ A_p^* = \frac{A_p - \mu_0(A_p)}{\sigma_0(A_p)} \]

For the case of unknown peak, first estimate the group \( p^* \) such that

\[ Z_{t^*}^* = \max \{ Z_t^* , t = 1, \ldots, k \} \text{ where } Z_t^* = \frac{Z_t - \mu_0(Z_t)}{\sigma_0(Z_t)} \quad t = 1, \ldots, k \]

Reject \( H_0 \) for large values of

\[ A_{p^*}^* = \frac{A_{p^*} - \mu_0(A_{p^*})}{\sigma_0(A_{p^*})} \]

The statistic \( A_{p^*}^* \) has an asymptotic null distribution that is standard normal. Thus, the test based on \( A_{p^*}^* \) is asymptotically distribution-free under null hypothesis. A Monte Carlo study demonstrated the fact that the modified tests are exactly distribution-free when the distributions are identical and the powers are higher than those of the original test (Mack and Wolfe, 1981) even though for small sample sizes the powers of the modified test are slightly lower than in the original tests.

**Chen and Wolfe’s test 2**

When the peak (\( p \)) of the umbrella is known à priori, Chen and Wolfe (1990b) generalized Chacko’s statistic to obtain a test for umbrella alternatives by

\[ \chi^2_{[p]} = \frac{12}{(N+1)} \sum_{i=1}^{k} \lambda_i (R_i^* - \frac{(N+1)}{2})^2 \]

where \( R_1^* \leq \ldots \leq R_p^* \geq \ldots \geq R_k^* \) is the isotonic regression of \( \bar{R}_1, \ldots, \bar{R}_k \) with weights \( \lambda_1, \ldots, \lambda_k \). We want to minimize

\[ \sum_{i=1}^{k} \lambda_i (r_i - \bar{R}_i)^2 \text{ subject to the constraints } r_1 \leq \ldots \leq r_p \geq \ldots \geq r_k \text{ and } \]

\[ \sum_{i=1}^{k} \lambda_i r_i = \frac{(N+1)}{2} \quad (8) \]

The following algorithm can be applied to obtain the isotonic regression \( R_1^* \leq \ldots \leq R_p^* \geq \ldots \geq R_k^* \). If \( R_1^* \leq \ldots \leq R_p^* \geq \ldots \geq R_k^* \) then \( R_i^* = \bar{R}_i \) (\( i = 1, \ldots, k \)). Otherwise start with \( \bar{R}_p \) the average rank of the peak group. Look for violators, where \( \bar{R}_i \) is a violator if \( \bar{R}_i > \bar{R}_{i+1} \) for \( i = 1, \ldots, p-1 \) or \( \bar{R}_i > \bar{R}_{i-1} \) for \( i = p+1, \ldots, k \).

Start the algorithm by choosing a violator and pooling it with its preceding average rank to form a weighted average rank. The violator and its preceding average rank will then be replaced by the weighted average rank. The weighted average rank compared with the adjacent rank average and so on. This procedure is continued until a set of quantities satisfying (8) is obtained. Chen and Wolfe (1990b) proved that for \( p = 1, \ldots, k \)
\[(\chi_{(p)}^2)^{1/2} = \max \left( \left\{ \frac{12}{N+1} \sum_{i=1}^{k} c_i \hat{\lambda}_i R_i \right\} \right) \]

where the maximum is taken over selections of \(c_1, \ldots, c_k\) such that

\[
\sum_{i=1}^{k} \lambda_i c_i = 0, \quad \sum_{i=1}^{k} \lambda_i c_i^2 = 1 \quad \text{and} \quad c_1 \leq \ldots \leq c_p \geq \ldots \geq c_k
\]

**Hettmansperger and Norton’s test**

Hettmansperger and Norton (1987) proposed two non-parametric procedures for testing null hypothesis against the umbrella alternatives for the case of known and unknown peak. For the case of known umbrella peak \(p\) and equally spaced treatments which correspond to \(\mu_i = \mu_0 + i\mu\) for \(i = 1, \ldots, p\) and \(\mu_i = \mu_0 + (2p-i)\) for \(i = p+1, \ldots, k\) they proposed rejecting \(H_0\) for large values of the statistic

\[
V_p = \left( \frac{12}{N+1} \right)^{1/2} \left\{ \sum_{i=1}^{p} \lambda_i (i - \bar{c}_w) \bar{R}_i + \sum_{i=p+1}^{k} \lambda_i (2p - i - \bar{c}_w) \bar{R}_i \right\}^{1/2}
\]

\[
\left( \sum_{i=1}^{p} \lambda_i (i - \bar{c}_w)^2 + \sum_{i=p+1}^{k} \lambda_i (2p - i - \bar{c}_w)^2 \right)^{1/2} \quad (9)
\]

where \(\bar{c}_w = \sum_{i=1}^{p} i \hat{\lambda}_i + \sum_{i=p+1}^{k} (2p-i) \hat{\lambda}_i\)

For description of \(\lambda_i, c_i\) and \(\bar{R}_i\) (\(i = 1, \ldots, k\)), see Chen and Wolfe’s test 2. For the unknown umbrella peak \(p\), (equally spaced alternative) they proposed rejecting \(H_0\) for large values of

\[
V^*_\text{max} = \max_{1 \leq t \leq k} V_t \quad \text{where } V_t \text{ is given by (9) for } t = 1, \ldots, k.
\]

**Simpson and Margolin’s test**

Simpson and Margolin (1986) investigated the relationship between an increasing dose-response when there is a potential for a drop in response at high doses. They suggested a recursive procedure using the non-parametric test for monotone trend proposed by Jonkheere and Terpstra. Set

\[
Q_j = \sum_{i=1}^{j-1} U_{ij} \quad \text{for } j = 2, \ldots, k \text{ where } U_{ij} \text{ is the Mann-whitney statistic.}
\]

Let
$S_t = \sum_{i=1}^{t-1} \sum_{j=t}^{t} U_{ij}$

be the Jonkheere-Terpstra statistic for the first $t$
samples ($t = 2, \ldots, k$).
Set
\[ M = \max_{2 \leq j \leq k} \left\{ j : Q_j \geq \frac{(n_j + ... + n_{j-1})n_j}{2} \right\} \]

The Simpson-Margolin test rejects \( H_0 \) for large values of
\[ S_M \left( \frac{1}{2} \right) = Q_2 + ... + Q_M \]

Rom et al. (1994) concluded that the Simpson-Margolin test has good power under the down-turn alternative and is a useful procedure when normality cannot be assumed.

**Chen and Wolfe’s test 3**

Chen and Wolfe (1993) developed a non-parametric distribution-free procedure to test whether there is at least one treatment that is better than the control when prior information about the umbrella pattern is available.

Suppose that \( X_{i1}, ..., X_{in_i} \) (i = 0, 1, ..., k) are \( k+1 \) independent random samples from population with continuous distribution functions \( F_i(x) = F(x - \mu_i) \) \( i = 0, 1, ..., k \) respectively. As usual the zero populations, \( i = 0 \), is the control group and the other \( k \) populations are treatment groups. One is interested in testing the null hypothesis that no treatment has any effect against \( \mu_i > \mu_0 \) for at least one \( i \) under the assumption that the treatments follow an umbrella pattern. Chen and Wolfe (1993) proposed one distribution-free test for comparing umbrella pattern treatment effects with a control when the peak of the umbrella is known a priori and two distribution-free tests in the case of unknown peak. They also suggested a way to estimate the lowest dose that is more effective than the control.

**Test procedure in the case of known umbrella peak**

Let \( R_{ij} \) be the rank of \( X_{ij} \) among the \( N = \sum n_i \) observations and let \( \bar{R}_i = \frac{\sum R_{ij}}{n_i} \) be the average rank of the \( i \)th sample, \( i = 0, 1, ..., k \). Suppose that the peak of the umbrella is known to be at the group \( p \) \((1 \leq p \leq k)\). Assume that \( n_0 \) is sample size of the control group and that all the treatment groups have same sample size, \( n_1 = ... = n_k = n \). Let \( R_{1*} \leq ... \leq R_{k*} \) be the isotonic regression of \( \bar{R}_1, ..., \bar{R}_k \) under the restriction \( \mu_1 \leq ... \leq \mu_p \geq ... \geq \mu_k \) for some \( p \). The algorithm how to obtain the \( R_{i*} \) is discussed on page 15. Reject the null hypothesis for large value of
\[ T_p = \frac{R_{p*} - \bar{R}_0}{\sqrt{\left\{ \frac{N(N+1)}{12} \right\} \left\{ \frac{1}{n+1/\mu_0} \right\}^{1/2}}} \]

Note that \( R_{p*} = \max_{1 \leq u \leq p \leq v \leq k} \sum_{i=u}^{v} R_i / (v-u+1) \)

Shirley’s (1977) test for comparing ordered treatment effects with a control is equivalent to this test based on \( T_k \). Suppose that \( N \to \infty \) in such a way that \( n/(n+n_0) \to \rho \) with \( 0 < \rho < 1 \). Then the statistic \( T_p \) converges in distribution to the statistic
\[ Y_p = \max_{1 \leq u \leq p, 1 \leq v \leq k} \sum_{i=u}^{k} w_i / (v - u + 1) \quad \text{as } N \to \infty \]

where the random vector \((w_1, \ldots, w_k)\) has a multivariate normal distribution with \(E(w_i) = 0, \text{Var}(w_i) = 1\) and \(\text{Cov}(w_i, w_j) = \rho, i \neq j = 1, \ldots, k\).

If the test rejects the null hypothesis, one can determine which dosage levels are more effective than the control by letting \(t_p(\alpha; n, n_0, k)\) be the value such that

\[
\Pr\{T_p \geq t_p(\alpha; n, n_0, k) \mid H_0\} = \alpha.
\]

Make the decision that \(\mu_i > \mu_0\) for \(1 \leq u \leq p \leq v \leq k\) if

\[
R_{u^*} - \bar{R}_0 \geq t_p(\alpha; n, n_0, k)[\{N(N+1)/12\}(1/n + 1/n_0)]^{1/2}
\]

and

\[
R_{v^*} - \bar{R}_0 \geq t_p(\alpha; n, n_0, k)[\{N(N+1)/12\}(1/n + 1/n_0)]^{1/2}
\]

If ties are present in the rankings one can modify the test based on \(T_p\) by replacing \(N(N+1)/12\) with

\[
N(N+1)/12 - \sum_{g \in G} (t_g^3 - t_g) / 12(N-1)
\]

where \(G\) is the set of group ties and \(t_g\) is the number of observations tied in the \(g\)th group.

**Test procedure in the case of unknown umbrella peak**

If prior information suggest that the peak group of the umbrella is relatively close to the \(k\)th group one can estimate the unknown peak, \(p_{s*}\) by a method suggested by Simpson and Margolin (1986).

Let \(U_{ij}\) be the Mann-Whitney statistic corresponding to the number of observations in the sample \(j\) that exceed observations in sample \(i\) and let

\[
Q_j = \sum_{i=1}^{j-1} U_{ij} \quad j = 2, \ldots, k. \quad \text{Let } p_{s*} = \max_{2 \leq j \leq k} \{j: Q_j \geq (j-1)n^2 / 2\}.
\]

Reject \(H_0\) for large values of

\[
T_{p_{s*}} = \max_{1 \leq u \leq p_{s*}, 1 \leq v \leq k} \left\{ \sum_{i=u}^{k} (\bar{R}_i - \bar{R}_0) / (v - u + 1) \right\} \left[ \{N(N+1)/12\}(1/n + 1/n_0) \right]^{1/2}
\]

If no information about the location of the peak group is available let

\[
Z_j = \sum_{i=1}^{j-1} U_{ij} + \sum_{i=j+1}^{k} U_{ij} \quad j = 1, \ldots, k
\]

First choose the group \(p_{m*}\) such that

\[
Z_{p_{m*}} = \max\{Z_j, j = 1, \ldots, k\}
\]
Reject the null hypothesis for large values of

\[ T_{ps} = \max_{1 \leq u \leq m, u \leq v \leq k} \left\{ \sum_{i=u}^{v} (\bar{R}_i - \overline{R}) / (v - u + 1) \right\} \left\{ N(N+1)/12 \right\}^{1/2} \left( 1/ n + 1/n_0 \right) \]

To determine which treatments are significantly better than the control let \( t_{ps}^\alpha (\alpha; n, n_0, k) \) and \( t_{pm}^\alpha (\alpha; n, n_0, k) \) be the upper \( \alpha \)th percentiles of the null distributions of \( T_{ps} \) and \( T_{pm} \) respectively. If the test rejects \( H_0 \), use the same procedure as in the case of known umbrella peak but substitute to the critical value \( t_{ps}^\alpha (\alpha; n, n_0, k) \) or \( t_{pm}^\alpha (\alpha; n, n_0, k) \).
**Test procedure** (Murkerjee, Robertson and Wright, 1987)

Assume equal sample sizes i.e. \( n_1 = n_2 = \ldots = n_k = n \).
Define weight vector \( \mathbf{w} = (w_0, w_1, \ldots, w_k) \) where \( w_0 = n_0 / n \) and \( w_i = n_i / n = 1 \) (\( i = 1, \ldots, k \)).
Also, define

\[
\mathbf{w} = (w_0, w_1, \ldots, w_k) \quad \text{where} \quad w_0 = n_0 / n \quad \text{and} \quad w_i = n_i / n = 1 \quad (i = 1, \ldots, k).
\]

and the cosine of the angle between \( \mathbf{x} \) and \( \mathbf{y} \) is \( \langle \mathbf{x}, \mathbf{y} \rangle/\| \mathbf{x} \| \| \mathbf{y} \| \). Let

\[
H_1' = H_1 \cap \{ \mathbf{x} \in \mathbb{R}^{k+1}. \sum_{i=0}^{k} w_i x_i = 0 \}
\]

\( H_1' \) is a closed, convex, pointed cone which is a subset of the \( k \)-dimensional subset of \( \mathbb{R}^{k+1} \) orthogonal to \( H_0 \). The cone \( H_1' \) is generated by non-negative multiples and convex combinations of the \( k \) corner vector:

\[
e_i = (e_{i0}, e_{i1}, \ldots, e_{ik}) = (-1, -1, \ldots, -1, w_0 + k - 1, -1, \ldots, -1) \quad (i = 1, \ldots, k)
\]

The angle between any two corners is \( \cos^{-1} \{-(w_0+k-1)^{-1}\} \). By symmetry the cone has a unique centre

\[
\mathbf{c} = \sum e_i = (-k, w_0, w_0, \ldots, w_0)
\]

which is in \( H_1' \) and makes equal angles with all of the corners.
The Abelson-Tukey contrast test modified by Schaafsma-Smid rejects \( H_0 \) for large values of

\[
T_A = \frac{\langle \mathbf{e}, \bar{X} \rangle}{S}.
\]

Under \( H_0 \), \( (n^{1/2} T_A)/\| \mathbf{c} \| \) has a Student-t distribution with \( m \) df.

Murkerjee, Robertson and Wright (1987) noted that this contrast test is very powerful when \( H_1 \) is the simple tree alternative and if prior knowledge is available that all of the differences \( \mu_i - \mu_0 \) are approximately equal. The contrast test has reasonable power when the \( H_1' \) cone is ”narrow” but the minimum power is very low when the cone \( H_1' \) is ”wide” for instance in the case of the simple tree. This test provides good power when the treatment means are approximately equal but is not recommended otherwise. Instead, Murkerjee, Robertson and Wright (1987) suggested a class of orthogonal contrast test based on the statistics

\[
T(r) = \max_{1 \leq i \leq k} \left( \frac{c_i(r), \bar{x}}{S} \right) \quad 0 \leq r \leq 1
\]

where

\[
c_i(r) = r \mathbf{e}_i + (1-r) \mathbf{e}_t \quad (i = 1, 2, \ldots, k)
\]
They proposed a class of tests given by (12) for a value of \( r = r_0 \), that makes the \( \{c_i(r)\} \) in (13) mutually orthogonal. The value of \( r_0 \) is given by

\[
r_0 = \begin{cases} 
  \left[2(w_0 + 1)\right]^{-1} & \text{if } (k - 2)w_0 = 1 \\
  -\left((w_0 + 1) + (w_0 + 1)^2 + w_0(k - 2) - 1\right)^{1/2} & \text{otherwise}
\end{cases}
\]

First compute \( r_0 \) and then compute for \( i = 1, 2, \ldots, k \) the vectors

\[
c_1' = r_0 e + (1-r_0)e_i \\
c_i = c_i' / \| c_i' \| \quad \text{using (10 and 11) and then}
\]

\[
T_i = \frac{(e, \bar{X})}{S} = \frac{\sum_{j=0}^{k} (c_{ij}, \bar{X}_j w_j)}{S} \quad (j = 0, \ldots, k; i = 1, 2, \ldots, k).
\]

Reject \( H_0 \) for large values of

\[
n^{1/2} T_o = n^{1/2} \max_{1 \leq i \leq k} T_i
\]

The random variables \( \{ST_i\} \) are independent, \( ST_i \) is distributed as \( N( (c_i, \mu), \sigma^2/n) \) and \( ST_i \) has zero mean if \( \mu \) satisfies \( H_0 \).

In general, Murkerjee et al. did not recommend the orthogonal contrast test due to the relatively low power, but this test is recommended in the case where one has no prior information about the relative effectiveness of the new drug due to its simplicity and uniform power characteristics. This test protects against all alternatives in \( H_1-H_0 \).

On the other hand if one is interested in testing whether the treatments are monotonically ordered and interested in finding the MED there are several contrast tests developed for this situation (Ruberg, 1989). Here we consider an experiment where \( k \) increasing dose levels are compared with a placebo control \( (i = 0) \). The observations \( y_{ij} \) are mutually independent \( (j = 1, 2, \ldots, n_i; i = 1, \ldots, k) \) where \( n_i \) is the number of experimental units at the \( i \)th dose level. Let \( \bar{y}_i \) be the sample means and let \( s^2 \) be an unbiased estimate of the common variance \( \sigma^2 \) based on \( v \) df where \( v = \sum_{i=0}^{k} n_i - (k + 1) \) df.

Here, only the case of equal sample sizes \( n_0 = n_1 = \ldots = n_k = n \) is considered.

The alternative hypothesis is now

\[
H_1 : \mu_0 = \mu_1 = \ldots = \mu_i < \mu_j \quad j = i+1, \ldots, k \text{ for any given } i < k.
\]

The \( \mu_i \) are not assumed to be monotonically ordered. Both Dunnett’s and William’s tests can be used for testing this hypothesis but Ruberg (1989) proposed three families of \( k \) contrasts for which the \( i \)th contrast in a given family is associated with identifying the \( i \)th as the MED. The three families of contrast are step contrast, an extension of the step contrast namely basin contrast and Helmert contrast.
Contrasts of the form

\[ a_{i0} \bar{y}_0 + a_{i1} \bar{y}_1 + \ldots + a_{ik} \bar{y}_k \]

where \( \sum_{j=0}^{k} a_{ij} = 0 \)

are used for testing \( H_0 \).

**Step contrast**

The coefficients for step contrast (Ruberg 1989) are given by

\[ a_{ij} = \begin{cases} 
-(k - i + 1) & j = 0, \ldots, i - 1 \\
1 & j = i, \ldots, k 
\end{cases} \]

The \( i \)th step contrast compares the average of the means of the highest \( k-i+1 \) dose levels with the average of all lower dose levels. The step contrasts are not orthogonal but they can be used to determine the MED by the following procedure proposed by Ruberg (1989).

**Test procedure**

If \( a_i \) (\( i = 1, 2, \ldots, k \)) are the vectors of contrast coefficients and \( \bar{y} \) is the random vector of treatment means, then under \( H_0 \) the random variable \((S_1, S_2, \ldots, S_k)\) where

\[ S_i = (a_i' \bar{y})/(a_i' a_i s^2/n)^{1/2} \]

has a \( k \)-variate t distribution with \( v \) df and a correlation matrix \( \{\rho_{ij}\} \) where \( \rho_{ij} = \text{corr}(S_i, S_j) \) (\( i \neq j \)).

Let \( S = \max_i \{S_i\} \) and let \( T_{k,v,(\rho_{ij})}^a \) be such that under \( H_0 \)

\[ \Pr(S > T_{k,v,(\rho_{ij})}^a) \leq \alpha. \]

If \( S = S_m > T_{k,v,(\rho_{ij})}^a \) then \( d_m \) is the MED.

In other words, the contrast that best matches the vector of observed means is used to identify MED.

**Basin Contrast**

The coefficients for basin contrasts (Ruberg 1989) are given by

\[ a_{ij} = \begin{cases} 
-(k - i + 1)(k - i + 2) / 2 & j = 0, \ldots, i - 1 \\
1 \cdot a_{i,j-1} + k + 1 & j = i, \ldots, k 
\end{cases} \]

The \( i \)th basin contrast compare the average of the means of the \( k-i+1 \) highest dose levels (\( 1 \leq i \leq k \)) with the average of the means of the zero dose level and the first \( i-1 \) dose levels. The means of the \( k-i+1 \) highest dose levels are weighted and the weights increase linearly with the dose level. This contrast is not orthogonal.
Test procedure

If $a_i$ (i = 1, 2, ..., k) are the vectors of contrast coefficients and $\bar{y}$ is the random vector of treatment means, then let $(B_1, B_2, ..., B_k)$ where

$$B_i = (a_i^T \bar{y})/(a_i^T a_i s^2/n)^{1/2}$$

which has a k-variate t-distribution with ν df and a correlation matrix $\{\rho_{ij}\}$. Let $B = \max_i\{B_i\}$. If $T^*_{k,\nu,(\rho_{ij})}$ is a critical value such that $Pr(B > T^*_{k,\nu,(\rho_{ij})}) \leq \alpha$ under $H_0$ and if $B = B_m > T^*_{k,\nu,(\rho_{ij})}$ then $d_m$ is the MED.

Tamhane et al. (1996) pointed out that both the step and basin contrasts have an excessive familywise error rate (FWE) which is defined

$$\text{FWE} = P\{\text{at least one } H_0 \text{ is rejected}\}.$$ 

Since in both cases the statistic has a non-central rather than a central t-distribution, the basin and step contrasts do not control the type 1 error rate and FWE and thus, tend to reject $H_0$ too often. Instead, Tamhane et al. proposed the Helmert, the linear and the pairwise contrasts for identifying the MED. In these cases, the general form of the t-statistic is given by

$$t_i = \frac{\sum_{j=0}^{k} a_{ij} y_j}{s \sqrt{\frac{a_{i0}^2}{n_0} + \sum_{j=1}^{n} a_{ij}^2 / n}} \quad 1 \leq i \leq k \quad (14)$$

The critical points depend on the joint distribution of the $t_i$ which is a multivariate t-distribution with ν df and a correlation matrix $\{\rho_{ij}\}$ where $\rho_{ij}$ is the correlation coefficient between the $i$th and $j$th contrasts $1 \leq i \neq j \leq k$.

Helmert contrast

The $i$th Helmert contrast (Ruberg 1989) is defined by

$$a_{ij} = \begin{cases} -1 & j = 0,1,...,i-1 \\ i & j = i \\ o & j = i+1,...,k \end{cases}$$

where $a_{ij}$ is the $j$th coefficient ($j = 0, 1, ..., k$) in the $i$th contrast ($i = 1, 2, ..., k$) The $i$th contrast compares the average of all lower dose level means with the $i$th dose level mean. Helmert’s contrasts are mutually orthogonal.
Test procedure (Ruberg, 1989)

If \(a_i\) (\(i = 1, 2, \ldots, k\)) is the vector of contrast coefficients and \(\bar{y}\) is the random vector of treatment means, then under \(H_0\)

\[H_p = (a_i' \bar{y})/(a_p' a_p s^2/n)^{1/2}\]

has a \(t\) distribution with \(v\) df.
Since the \(a_i\) are mutually orthogonal the \(\rho_{ij} = 0\) when \(n_0 = n\) and the critical value for each test, \(M^a_{k,v}\), is obtained from the Studentized maximum distribution (Hochberg and Tamhane, 1987).

If \(H_1 > M^a_{k,v}\) then \(\mu_1\) is the MED;
if \(H_1 < M^a_{k,v}\) and \(H_2 > M^a_{k,v}\) then \(\mu_2\) is the MED.
This procedure continues until
i. \(H_m > M^a_{k,v}\) for some \(1 \leq m \leq k\) where \(\mu_m\) is the MED or
ii. \(H_k < M^a_{k,v}\) where \(H_0\) cannot be rejected.

Linear contrasts
The general form of these linear contrast (Rom et al. 1994) is given by

\[a_{ij} = \begin{cases} -i & j = 0 \\ a_{i,j-1} + 2 & j = 1, \ldots, i \\ 0 & j = i + 1, \ldots, k \end{cases}\]

Pairwise contrasts
The \(i\)th pairwise contrast is \(y_i - y_0\) (\(1 \leq i \leq k\)). The \(t\)-statistic is given by

\[t_i = \frac{y_i - y_0}{s \sqrt{\frac{1}{n_0} + \frac{1}{n}}} \quad 1 \leq i \leq k\]

The correlation coefficients are given by \(\rho_{ij} = \frac{n}{n_0 + n}\).

Stepwise testing procedure
For the Helmert, linear and pairwise contrasts, Tamhane et al. (1996) described two types of stepwise testing procedure; step-down and step-up, for testing the null hypothesis against the alternative hypothesis i.e.

\[H_{0i} : \mu_0 = \mu_1 = \ldots = \mu_i\]
\[H_{1i} : \mu_0 = \mu_1 = \ldots = \mu_{i-1} < \mu_i \quad (1 \leq i \leq k)\]
The $i$th dose is identified as the MED and the $\mu_i$ are not assumed to be monotonically ordered.

Marcus et al. (1976) proposed a closed testing procedure that controls the FWE in this type of multiple hypothesis testing procedure. Firstly, one must form a closure of the family of hypotheses $\{H_{0i}; 1 \leq i \leq k\}$ for any set of indices $1 \leq i_1 \leq i_2 \leq \ldots \leq i_m \leq k$

$$H_{0i_1} \cap H_{0i_2} \cap \ldots \cap H_{0i_m} = H_{0i_m}.$$ Then, separate $\alpha$-level tests of the individual $H_{0i}$ are performed.

In the step down procedure, first order the $t$-statistics $t_1 \leq t_2 \leq \ldots \leq t_k$ and order the corresponding hypotheses $H_{0(1)} \leq H_{0(2)} \leq \ldots \leq H_{0(k)}$. Then a hypothesis $H_{0i}$ is rejected at level $\alpha$ iff all the hypotheses $H_{0j}$ are significant at level $\alpha$ for $j \geq i$. Thus, the procedure is performed in a step down manner.

The step up procedure proposed by Dunnett and Tamhane (1992) is based on a set of critical constants $c_1 \leq c_2 \leq \ldots \leq c_k$ (to determine $c$ see next page) which is compared with the corresponding t-statistic. The step up procedure start by testing the smallest t-statistic and work upward, accepting one hypothesis at a time. When $t_i \geq c$ the procedure is stopped by rejecting the hypothesis $H_{0(i)}, H_{0(i+1)}, \ldots, H_{0(k)}$.

**Stepdown procedure 1** (Marcus 1976)

At the $i$th step let $k_i$ be the number of hypotheses still to be tested. Relabel the order statistics $t_1 \leq t_2 \leq \ldots \leq t_{k_i}$ and the corresponding hypotheses as

$$H_{0(1)}, H_{0(2)}, \ldots, H_{0(k)}.$$ Test $H_{0(k_i)}$ by comparing $t_{k_i}$ with $c_{k_i} = t_{k_i, v, \rho}^\alpha$ (tables in Marcus, 1976)

Reject $H_{0(k_i)}$ if $t_{k_i} \geq t_{k_i, v, \rho}^\alpha$ and all hypotheses whose rejection is implied by it and go to the next step, otherwise stop testing.

When testing is stopped, estimate the MED as the minimum index of the rejected hypotheses.

**Step down procedure 2** (Rom et al. 1994)

This is a simpler and a more flexible procedure which controls the FWE and does not require ordering of the t-statistics and controls the FWE.

Reject $H_{0(i)}$ iff each $H_{0(j)}$ is significant for all $j \geq i$ using an $\alpha$-level t-test i.e. $t_j \geq c = t_v^\alpha$ where $t_v^\alpha$ is the upper critical point of Student’s t-distribution with $v$ df.

**Step up procedure 1** (Dunnett and Tamhane 1992)

Let $c_1 \leq c_2 \leq \ldots \leq c_k$ be the critical constants for step up for given $k$, $v$, $\alpha$, and $\rho$ (tables in Dunnett and Tamhane 1992).

Order the t-statistics; $t_1 \leq t_2 \leq \ldots \leq t_k$.

Test the hypothesis $H_{0(i)}$ iff $t_j < c_j$ for $j = 1, \ldots, i-1$. 
If \( t_i > c_i \) then proceed to test \( H_{0(i+1)} \) otherwise, stop testing and reject the hypotheses

\[
H_{0(i)}, H_{0(i+1)}, \ldots, H_{0(k)}
\]

and any hypotheses whose rejection is implied by them. Estimate the MED as

\[
MED^* = i^* = \min\{(i), \ldots, (k)\}.
\]

Dunnett and Tamhane showed that the step up procedure strongly controls the FWE.

**Step up procedure 2** (Tamhane et al 1996)

This procedure is based on unordered t-statistics and a common critical constant \( c \). Test \( H_{0i} \) iff \( t_j < c_j \) for \( j = 1, \ldots, i-1 \).

If \( t_i \geq c_i \) then stop testing and reject \( H_{0(i)}, H_{0(i+1)}, \ldots, H_{0(k)} \) otherwise go to the next step.

If no hypotheses are rejected then no dose is declared as MED, otherwise MED is estimated as

\[
MED^* = i^* = \min \{i: t_i \geq c\}.
\]

To determine \( c \), consider any true hypothesis \( H_{0(i)} \). Then

\[
\text{FWE} = P\{ \text{reject } H_{0i} \} = 1 - P\{t_1 < c, \ldots, t_i < c\}
\]

where \( t_1, \ldots, t_i \) have an i-variate t-distribution with \( v \) df and common correlation \( \rho \).

For determining \( c \), solve the equation

\[
P\{ t_1 < c, \ldots, t_k < c \} = 1 - \alpha
\]

where \( c = t_{\alpha, v, \rho}^k \).

Tamhane et al. (1996) concluded that the best procedures based on a power study were; step down 1 for Helmert’s contrast, step down 2 for linear contrast and also William’s test (see page 6). In general, step down procedures were preferred over step up procedures and if the dose-response function was highly non-monotone the step down 2 procedure should not be used.

The step down 1 and 2 procedures for linear contrast and Williams’ test were the best procedures in terms of bias when the dose-response function was monotone. For non-monotone functions the step down 1 procedure for linear, Helmert and pairwise contrasts were recommended.

In the case where the normal theory assumptions are not satisfied the only test recommended is the step down 2 procedure for pairwise contrast, but otherwise this test is low both in terms of power and bias.
Power comparisons and conclusions

In this section, the relative powers of some of the different test procedures reviewed in this paper will be compared. Not by any means, I have no intention to give a complete comparison between all test procedures, only a selection of the most commonly used procedures are compared.

The powers of Bartholomew’s test and Williams’ test were compared by Shirley (1979). It was found that Bartholomew’s test is more powerful than Williams’ test and therefore preferable. But Williams’ test is a good alternative since there are more tables of significant points available for Williams’ test and it is just as powerful as Bartholomew’s test at detecting low dose effects.

Shirley (1985) concluded that Jonkheere’s distribution-free and non-parametric test had higher type I error than both Bartholomew’s test and Williams’ test. Jonkheere’s test was not very powerful except when the responses are linear and it is recommended to be used to test hypotheses concerning ordered distributions. Compared with these three tests Dunnett’s test was slightly more powerful than Jonkheere’s test, but the type I error rate was unacceptably high. Also, Dunnett’s test should be used under normality assumptions but nothing is assumed about the response shape. In Williams’ test one assume that monotonicity of response is known to occur in a known direction. Williams (1971) found that his proposed test was superior to Dunnett’s test but slightly less powerful than Bartholomew’s test which becomes more powerful as k increases. This was also found by Marcus (1976).

Also, Marcus compared the Bartholomew’s test and Williams’ test with the Abelson-Tukey contrast test. The Abelson-Tukey contrast test is only recommended to be used if prior information about the ordering of the means is available. Murkerjee et al. (1987) found that Abelson-Tukey contrast test has good power when the treatment means are approximately equal but otherwise it is not recommended at all. Instead the orthogonal contrast test is recommended due to its uniform power characteristics. The basin and step contrasts proposed by Ruberg (1989) are not recommended at all due to their inability to control the FWE.

Tamhane et al. (1996) and Dunnett and Tamhane (1992) recommended the step down procedures over the step up procedures. Specially, the step down procedures for Helmert’s contrast and linear contrast were most powerful. Here, the treatment responses are assumed to be normally distributed but not monotonically ordered. When the normality assumptions are not satisfied, the step down procedure for pairwise contrasts is recommended.

The Jonkheere’s (1954) test, Chacko’s (1963) test, Mack and Wolfe test (1981) Hettmansperger and Norton’s (1987) test and Simpson-Margolin (1986) were compared in a power study by Chen and Wolfe (1990b). The Jonkheere’s test was found to be more powerful than Chacko’s test. When the umbrella alternative was considered, different conclusions were made depending on whether the peak of the umbrella was known or not. Mack and Wolfe (1981) recommended to use Jonkheere’s test in ordered settings and to use their proposed test in umbrella alternative settings, since their test has excellent power when used in the right situation.

If the peak is known both the Mack and Wolfe’s test and Hettmansperger and Norton’s test was better than Chacko’s test. For equal spacing alternatives, Hettmansperger and Norton’s test was more powerful than Mack and Wolfe’s test. If the peak is unknown the Simpson and Margolin test is a more powerful test than the other test.
but only in the case where the peak is close to the \( k \)th treatment group, otherwise this test has low power.

The modified Mack and Wolfe’s test (Chen and Wolfe, 1990a) has slightly lower power than the original Mack and Wolfe test for small sample sizes but instead the modified tests are distribution-free.

Finally, Chen and Wolfe (1993) proposed a new test statistics for umbrella alternatives. In general, the Mack and Wolfe (1981) test has higher power but in the case where one is confident about the location of the peak group, the proposed test statistic should be used.

As Shirley (1985) pointed out there is no test which can be used in all experimental situations for testing all different alternative hypotheses. Depending on the degree of knowledge, for instance, whether or not the researcher knows if the dose-responses are monotonically ordered or not, different tests should be used. As in the case of the rank tests for umbrella alternatives, there are different tests depending on if the peak of the umbrella is known or not. Also, the tests differ with respect of whether the data is normally distributed or not, and if the test is parametric or non-parametric and distribution-free or not. Therefore, no simple recommendation can be given which test is the best. It depends on the degree of knowledge about the shape of the responses and the aim of the study. Thus, to maximize the power it is worthwhile using more specialized tests when it is appropriate.

**Acknowledgements**

I would like to thank Staffan Uvell, PhD, Department of Mathematical Statistics, Umeå University, for his kind help in the conduct of this paper and for his valuable comments. I would also like to thank all employees at the Department of Mathematical Statistics for all support and encouragement during my studies at Umeå University.
Bibliography


Williams, D. A. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics, 27, 103-117.

Bibliography


Rom, D. M., Costello, R. J. and Connell, L.T. (1994). On closed test procedures for dose- 
Assoc., 84, 816-822.
Schaafsma, W. And Smid, L. J. (1966). Most stringent somewhere most powerful tests against 
Simpson, D. G. and Margolin, B. H. (1986). Recursive nonparametric testing for dose- 
response relationships subject to downturns at high doses. Biometrika, 73, 589-596.
Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall’s test against 
Mathem., 55, 327-333.
Williams, D. A. (1971). A test for differences between treatment means when several dose 
levels are compared with a zero dose control. Biometrics, 27, 103-117.
Williams, D. A. (1972). The comparison of several dose levels with a zero dose control. 
Biometrics, 28, 519-531.