

# Nonparametric comparison of several mean values with mild adaptation to some sample characteristics

**Citation for published version (APA):**

Dijkstra, J. B. (1984). *Nonparametric comparison of several mean values with mild adaptation to some sample characteristics*. (Computing centre note; Vol. 20). Technische Hogeschool Eindhoven.

**Document status and date:**

Published: 01/01/1984

**Document Version:**

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

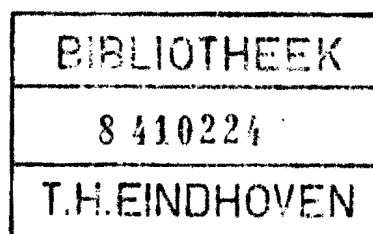
providing details and we will investigate your claim.

THE-RC 55041

Eindhoven University of Technology  
Computing Centre Note 20

Nonparametric comparison of several mean  
values with mild adaptation to some sample  
characteristics

Jan B. Dijkstra



Prepared for the conference on Computational Statistics (COMPSTAT).

August 27-31, 1984

Prague, Czechoslovakia.

NONPARAMETRIC COMPARISON OF SEVERAL MEAN VALUES WITH MILD ADAPTATION TO  
SOME SAMPLE CHARACTERISTICS

Jan B. Dijkstra

Eindhoven University of Technology  
The Netherlands

Keywords: Kruskal & Wallis, Mood & Brown and Van der Waerden test, logistic, double exponential and normal distribution, adaptation, asymptotic relative efficiency.

ABSTRACT

The rank tests by Kruskal & Wallis (KW), Mood & Brown (MB) and Van der Waerden (VdW) are asymptotically optimal for logistic (L), double exponential (DE) and normal (N) distributions respectively. These tests belong to an unlimited family and it is possible to look at the data and then to construct an optimal member of this family. This is "strong adaptation" and it only works for very large samples. One can also decide which of the three distributions (L, DE or N) gives optimal fit for the data and then apply KW, MB or VdW respectively. One might expect something to be gained by this form of "mild adaptation" because some of the asymptotic relative efficiencies of KW/MB, KW/VdW and MB/VdW differ seriously from 1 or L, DE and N distributions. Several criteria are considered for choosing between the tests, (KW, MB or VdW) given some sample characteristics. It is well known that in this situation one should consider the combined sample and ignore the origin of the individual values. For some of the above mentioned criteria this principle is seriously violated, so that the resulting tests do not remain permutation tests. A simulation study is performed in order to find out how much this violation affects the control over the chosen size  $\alpha$ . Powers are estimated for various alternatives, also by using simulation techniques. The conclusion: "Mild adaptation is possible in this situation without serious harm in the sense of size-control. Some extra power can be gained by it, though not for small samples".

INTRODUCTION

Let  $x_1, \dots, x_n$  be a combination of  $k$  samples.  $S_j$  denotes the collection of indices in sample  $j$  and  $n_j$  the corresponding sample size.  $R_i$  is the rank of observation  $x_i$ . Now the test statistic of Kruskal & Wallis (1952) is given by:

$$Q_1 = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{1}{n_j} \left[ \sum_{i \in S_j} R_i \right]^2 - 3(N+1) \quad (1)$$

The notation is from Hajek and Sidak (1967). The test statistic of Mood and Brown (1950) is:

$$Q_2 = 4 \sum_{j=1}^k \frac{1}{n_j} A_j^2 - N, \text{ where} \quad (2)$$

$$A_j = \sum_{i \in S_j} \frac{1}{2} [\text{sign}(R_i - \frac{1}{2}(N+1)) + 1]. \quad (3)$$

The third test in this study is from Van der Waerden. It uses the standard normal distribution function  $\Phi$ . The test statistic is given by:

$$Q_3 = \frac{N-1}{h} \sum_{j=1}^k \frac{1}{n_j} \left[ \sum_{i \in S_j} \Phi^{-1} \left( \frac{R_i}{N+1} \right) \right]^2, \text{ where} \quad (4)$$

$$h = \sum_{i=1}^N \left[ \Phi^{-1} \left( \frac{i}{N+1} \right) \right]^2 \quad (5)$$

These methods test the equality of  $k$  location parameters for distributions that are at least similar in shape and scale. Although this is the only thing one can achieve with these tests, the formal hypothesis  $H_0$  is "all samples come from the same distribution". Another similarity between these tests is their behaviour under  $H_0$ . In that situation  $Q_1, Q_2$  and  $Q_3$  are asymptotically distributed as  $\chi_{k-1}^2$ .

TWO STRATEGIES IN ADAPTATION

The asymptotic behaviour of the three tests if  $H_0$  does not hold can differ considerably. For each test a distribution exists for which the power is asymptotically optimal. These distributions are:

test	distribution
Kruskal & Wallis	logistic
Mood & Brown	double exponential
Van der Waerden	normal

table 1: asymptotic optimality

It is possible to give a general form for the kind of rank tests mentioned in this paper:

$$Q = \frac{N-1}{h^*} \left[ \sum_{j=1}^k \frac{1}{n_j} T_j^2 - N\bar{a}_N^2 \right], \text{ where} \quad (6)$$

$$h^* = \sum_{i=1}^N [a_N(i) - \bar{a}_N]^2 \text{ and} \quad (7)$$

$$T_j = \sum_{i \in S_j} a_N(R_i) \quad (8)$$

The  $a_N(i)$  are scores that can be chosen in order to get optimal power for a certain distribution. It is possible to look at the original sample and use it to estimate the underlying distribution. This way one might hope to achieve asymptotic optimality of the test. Unfortunately one needs very big samples to have any success, and therefore this approach of "strong adaptation" has already been rejected in the literature [Huber (1972)].

An alternative and less ambitious approach is the following. Look at the data and determine what the distribution looks like (logistic, double exponential or normal). Then select the appropriate test from table 1. Please note that the actual distribution can differ greatly from these three, but there will always be one with a minimal difference for a chosen criterion. This kind of "mild adaptation" will only be worth the trouble if the powers of the tests differ enough for the above mentioned distributions. The next section is included to give an impression of what one might hope to achieve.

#### ASYMPTOTIC RELATIVE EFFICIENCY

A well known criterion for comparing the powers of two tests is the asymptotic relative efficiency ARE. Let A and B be tests and let  $a$  and  $b$  be the corresponding numbers of observations involved. For a chosen size  $\alpha$  both tests are used for the same  $H_0$  against a class of alternatives  $\{H_n \mid n \in N\}$ . Now  $ARE_{A,B}$  is defined as the asymptotic value of  $b/a$  when  $a$  varies such that the powers are (and remain) equal while  $b \rightarrow \infty$  and  $H_n \rightarrow H_0$ .

Andrews (1954) gives a formula for the ARE of the Mood & Brown test relative to the Kruskal & Wallis test.

$$ARE_{MB, KW} = 1/3 \left[ \frac{F'(M)}{\int_{-\infty}^{\infty} F'(x) dF(x)} \right]^2 \quad (9)$$

where  $M$  is the median of  $F$ . Unfortunately no formula like this seems to be published where MB or KW is replaced by VdW, but the same result can be attained by another approach. The family given by (6), (7) and (8) contains a member described by Terry and Hoeffding (1960) that is very similar to the Van der Waerden test and that has the same ARE's [Bradley (1968)]. Hodges and Lehmann (1961) examined the two-sample situation for  $ARE_{TH, W}$  ( $W$  stands for Wilcoxon which is KW with  $k = 2$ ). Using these results and formula (9) one can construct the following table.

distribution	$ARE_{VdW, KW}$	$ARE_{VdW, MB}$	$ARE_{KW, MB}$
normal	$\pi/3$	$\pi/2$	$3/2$
logistic	$3/\pi$	$4/\pi$	$4/3$
double exponential	$8/3\pi$	$2/\pi$	$3/4$

table 2: asymptotic relative efficiency

Looking at this table one can conclude that mild adaptation as described in the previous section is worth trying. Especially if it is not known what kind of distribution one is dealing with.

#### CRITERIA FOR SELECTING THE TEST

In a simulation study four samples were generated from normal [Box & Muller (1958)], logistic [Newman & Odell (1971)] or double exponential [V Putten & Vd Tweel (1979)] distributions. The samples contained 15 observations. The location parameter was set at four different values, so the combined sample did not look like the separate samples that came from the chosen distribution. Now this is a bit of a problem. Hajek (1969) remarks that if one wants to select a rank test on the basis of some sample characteristics, one should ignore to which samples the individual values belong.

Each experiment was carried out a number of times, where the four samples came from the same (but shifted) distribution. Several ideas were tried and the quality of an idea is indicated by the number of times that the correct distribution was selected.

First idea: The sample kurtosis of the combined sample.

$$K = \left[ \sum_{i=1}^N (x_i - \bar{x})^4 / N \right] / \left[ \sum_{i=1}^N (x_i - \bar{x})^2 / N \right]^2 - 3 \quad (10)$$

The kurtoses for the distributions under consideration are well known:

distribution	K	criterion
normal	0	0.6
logistic	1.2	2.1
double exponential	3	

table 3: kurtosis

The values are reasonably distinct which was to be expected since the major difference between the distributions from the table lies in their tail length. The kurtosis can be used as follows:

kurtosis	test
$K < 0.6$	Van der Waerden
$0.6 < K < 2.1$	Kruskal & Wallis
$K > 2.1$	Mood & Brown

table 4: selection on K

If one uses the combined sample, the kurtosis will be incorrectly estimated if  $H_0$  does not hold. The simulation study confirmed this. So now comes the:

Second idea: Make a copy of the observations so that every sample is shifted to make the means equal. Then compute K for the combined sample and proceed as with the first idea. Please note that the principle mentioned by Hajek is violated with this approach. Some other ideas will have the same problem, that will be dealt with later in this paper.

The number of correct selections improved very much, but was still not satisfactory.



Third idea: Centralize on the medians instead of the means. Since the experiment involves the double exponential distribution with very heavy tails this is a natural thing to try. The attempt yielded a slight improvement but not enough.

Fourth idea: Compute  $K_i$  ( $i = 1, \dots, k$ ) and use  $K^* = \frac{\sum_{i=1}^k n_i K_i}{N}$  (11)

The results were similar to centralization on the means, which is just what one might expect. These four disappointments involving the kurtosis lead to the conclusion that another criterion should be considered.

Fifth idea: Let  $U_\beta$  be the sum of the upper  $N \cdot \beta$  observations for  $0 < \beta < 1$ . If  $N \cdot \beta$  is not an integer then one observation should only be fractionally included.  $L_\beta$  has a similar meaning, where  $L$  stands for lower. Using these concepts Hogg (1974) suggests as a measure for the tail length:

$$Q^* = \frac{10(U_{0.05} - L_{0.05})}{U_{0.5} - L_{0.5}} \quad (12)$$

The experiments with  $K$  make it natural to try  $Q^*$  only with centralization on the medians. The use of  $Q^*$  as a criterion for selection is given in the following tables.

distribution	$Q^*$	criterion
normal	2.58	2.71
logistic	2.85	3.07
double exponential	3.30	

table 5: criterion  $Q^*$

critterion	test
$Q^* < 2.71$	Van der Waerden
$2.71 < Q^* < 3.07$	Kruskal & Wallis
$Q^* > 3.07$	Mood & Brown

table 6: selection on  $Q^*$ 

The use of  $Q^*$  improved the probability of a correct selection enormously. The number of times that the simulated distribution was recognized was highly satisfactory and so there seems to be no reason to look any further.

#### INCORRECT USE OF INFORMATION

The aim of this paper is to produce an adaptive test that has for a large family of distributions more power than the separate tests already mentioned. The adaptation lies in the selection of the test on the value of  $Q^*$  that is estimated from the combined sample after centralisation on the medians. If this selection is based on an incorrect use of information, there is some danger that the following situation will be met:

$$P\{\text{rejection } H_0\} > \alpha \quad (13)$$

Suppose that the three separate tests were applied with the decision-rule: reject  $H_0$  if  $\max(Q_1, Q_2, Q_3) > \chi_{k-1}^2(\alpha)$ . In this case one does not need a simulation study to find out that (13) is just what will happen.

An adaptive test can never be a rank test in the proper sense. However, it would remain a permutation test if the computation of  $Q^*$  were not preceded by centralisation on the medians. But without this centralization it can easily happen if  $H_0$  does not hold that a normal distribution will be classified as a double exponential distribution, so that the procedure chooses the worst test with an enormous loss of power relative to the correct selection (see table 2).

The worst thing that can happen as a consequence of the centralisation is given in (13). Since no analytical approach seems to exist, a simulation study was performed in order to check the control over the chosen size  $\alpha$ . Under  $H_0$  it will have asymptotically no effect to centralise since in that case the medians will be equal. So the danger lies in the smaller samples.

For very small samples the use of  $\chi^2$  is dubious, since the test statistics have this distribution only asymptotically. If some  $n_i$  is less than 7 it seems better to use the  $\beta_q^p$  approximation by Wallace (1959) where  $p$  and  $q$  are functions of  $(n_1, \dots, n_k)$ .

So if problem (13) arises one can only expect to find it for moderate samples, because for smaller samples it will be overshadowed by the discrepancy between the test statistic and the  $\chi^2$  distribution, and for bigger samples it will asymptotically disappear. These considerations suggest that  $n_i = 15$  for  $i = 1, \dots, 4$  is a sensitive situation. A simulation study under  $H_0$  for normal, logistic and double exponential distributions showed not the slightest tendency to problem (13).

With a mixture of 1500 replications from each of these distributions the fraction of rejections was 0.0497 for  $\alpha = 0.005$ . Therefore the adaptive test with  $Q^*$  computed after centralisation on the medians seems acceptable in the sense of size-control.

#### A COMPARISON OF POWERS

In a simulation study samples were generated from normal, logistic and double exponential distributions. The adaptive test, as well as the tests of Van der Waerden, Kruskal & Wallis and Mood & Brown were applied for various patterns concerning the location parameters (see table 10). For every situation equal sample sizes of 15 and 40 elements were generated. Powers were estimated by the fraction of rejections for 300 replications. The results are given in tables 7 to 9.

location	$n_1$	adaptive	VdW	KW	MB
A	15	0.37	0.36	0.34	0.20
	40	0.80	0.79	0.78	0.59
B	15	0.45	0.44	0.44	0.28
	40	0.93	0.92	0.91	0.76
C	15	0.65	0.65	0.65	0.43
	40	0.99	0.99	0.99	0.94
D	15	0.74	0.74	0.73	0.48
	40	0.99	0.99	0.99	0.93

table 7: normal distribution

location	$n_1$	adaptive	VdW	KW	MB
E	15	0.37	0.36	0.37	0.27
	40	0.80	0.76	0.80	0.65
F	15	0.37	0.38	0.39	0.28
	40	0.90	0.89	0.91	0.80
G	15	0.88	0.90	0.90	0.73
	40	1.00	1.00	1.00	1.00
H	15	0.90	0.91	0.92	0.78
	40	1.00	1.00	1.00	1.00

table 8: logistic distribution

location	$n_1$	adaptive	VdW	KW	MB
I	15	0.28	0.27	0.32	0.26
	40	0.76	0.72	0.76	0.76
J	15	0.35	0.31	0.37	0.31
	40	0.86	0.81	0.86	0.86
K	15	0.80	0.83	0.86	0.75
	40	1.00	1.00	1.00	1.00
L	15	0.83	0.85	0.88	0.77
	40	1.00	1.00	1.00	1.00

table 9: double exponential distribution

location	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$
A	0.00	0.10	0.20	0.70
B	0.00	0.05	0.30	0.80
C	0.00	0.15	0.30	1.05
D	0.00	0.10	0.45	1.05
E	0.00	0.20	0.40	1.20
F	0.00	0.10	0.50	1.30
G	0.00	0.40	0.80	2.40
H	0.00	0.20	1.00	2.40
I	0.00	0.10	0.20	0.80
J	0.00	0.10	0.35	0.90
K	0.00	0.20	0.40	1.60
L	0.00	0.20	0.70	1.60

table 10: location parameters

In order to quantify the comparative merits of these tests the number of rejections for the combined simulation study involving 7200 situations was computed for every test. This resulted in the following table after linear scaling on the interval from zero to one for both sample sizes:

test	$n_i = 15$	$n_i = 40$
adaptive	0.89	1.00
Van der Waerden	0.90	0.78
Kruskal & Wallis	1.00	0.94
Mood & Brown	0.00	0.00

table 11: comparative powers

For  $n_i = 15$  the adaptive test is not the most powerful one. The reason for this disappointment can be found in table 9. If one compares the tests by Kruskal & Wallis and Mood & Brown for the double exponential distribution, it can be seen that the first one is consistently the best for  $n_i = 15$ . For  $n_i = 40$  their powers are equal, so that the asymptotic superiority of the Mood & Brown test for this distribution can only be expected to show itself for  $n_i > 40$ .

So for  $n_i = 15$  the correct recognition of a double exponential distribution leads to a loss of power. And for  $n_i = 40$  it does not matter if the test discriminates between the logistic and double exponential distribution since for that sample size there seems to be no difference in behaviour between the separate tests involved.

The superiority of the adaptive test for  $n_i = 40$  comes only from its ability to recognize normal distributions. A simpler method that uses the Van der Waerden test for  $Q^* < 2.71$  and the Kruskal & Wallis test for  $Q^* > 2.71$  would have the same power for this sample size.

There is no doubt that the asymptotic efficiency of the adaptive test relative to the separate tests for a mixture of normal, logistic and double exponential distributions with equal occurrences is consistently greater than one (see table 2). Finite samples, however, should contain more than 40 observations if any profit from the rather complex procedure is to be derived.

Such sample sizes are very rare in practice. And so the adaptive test is of a very limited value.

ACKNOWLEDGEMENTS

Prof.dr. R. Doornbos and Dr. W. Albers were willing to discuss the set-up of this experiment, before and during its development. I like to thank them for their helpful suggestions.

T. Smeulders and R.V.H. Rooijackers wrote the simulation program that produces the tables included in this paper. The enthusiasm and thoroughness of these two students was essential for its development.

Dr. H.N. Linssen read an earlier version of this paper. I feel sure that his comments have improved its quality.

REFERENCES

- [1] Kruskal, W.H. and Wallis, W.A. (1952)  
Use of ranks in one-criterion variance analysis  
JASA 47, 583-621. Errata: JASA 48 (1953), 910.
- [2] Hajek, J. and Sidak, Z. (1967)  
Theory of rank tests  
Academic Press, New York.
- [3] Brown, G.W. and Mood, A.M. (1950)  
On median tests for linear hypotheses  
Proc. 2nd. Berkeley Symposium, 159-166.
- [4] Huber, P.J. (1972)  
Robust statistics: a review  
The Annals of Mathematical Statistics, Vol. 43, No. 4, 1041-1067.
- [5] Andrews, F.C. (1954)  
Asymptotic behaviour of some rank tests for analysis of variance  
The Annals of Mathematical Statistics 25, 724-736.
- [6] Terry, M.E. (1960)  
An optimum replicated two-sample test using ranks  
Contributions to Probability and Statistics, Stanford University Press, 444-447.
- [7] Bradley, J.V. (1968)  
Distribution-free statistical tests  
Prentice-Hall, Englewood Cliffs.

- [8] Hodges, J.L. and Lehmann, E.L. (1961)  
Comparison of the Normal Scores and Wilcoxon tests  
Proceedings of the Fourth Berkely Symposium on Mathematical  
Statistics and Probabilitiy, University of California Press, Vol. 1,  
307-317.
- [9] Box, G.E.P. and Muller, M.E. (1958)  
A note on the generation of random normal deviates  
The Annals of Mathematical Statistics, 29.
- [10] Newman, T.G. and Odell, P.L. (1971)  
The generation of random variates  
Griffin, London.
- [11] Putten, C. van and Tweel, I. van der (1979)  
On generating random variables  
Mathematisch Centrum, Amsterdam.
- [12] Hajek, J. (1969)  
A course in Nonparametric Statistics  
Holden-Day, San Francisco
- [13] Hogg, R.V. (1974)  
Adaptive Robust Procedures: A Partial Review and some suggestions  
for future applications and theory  
JASA, Vol. 69, 909-923.
- [14] Wallace, D.L. (1959)  
Simplified Beta-approximations to the Kruskal-Wallis H Test  
JASA 54, 225-230.