

Robustness of multiple comparisons against variance heterogeneity

Citation for published version (APA):

Dijkstra, J. B. (1983). *Robustness of multiple comparisons against variance heterogeneity*. (Computing centre note; Vol. 17). Technische Hogeschool Eindhoven.

Document status and date:

Published: 01/01/1983

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

Take down policy

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

openaccess@tue.nl

providing details and we will investigate your claim.

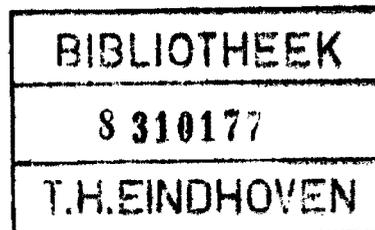


Bibliotheek

Eindhoven University of Technology
Computing Centre Note 17

Robustness of Multiple Comparisons
against variance heterogeneity

Jan B. Dijkstra



Prepared for the Conference on Robustness of
Statistical Methods and Nonparametric Statistics.

May 29 to June 4, 1983
Schwerin, GDR.

ROBUSTNESS OF MULTIPLE COMPARISONS AGAINST VARIANCE HETEROGENEITY

Jan B. Dijkstra

Computing Centre, Eindhoven University of Technology.

ABSTRACT

If $H_0: \mu_1 = \dots = \mu_k$ is rejected for normal populations with classical one way analysis of variance, it is usually of interest to know where the differences may be. If the population variances are equal there are several approaches one might consider:

1. Least Significant Difference test (Fisher, 1935)
2. Multiple Range test for equal sample sizes (Newman, 1939)
3. An adaptation for unequal sample sizes (Kramer, 1956)
4. Multiple F-test (Duncan, 1951)
5. Multiple Comparisons test (Duncan, 1952).

For all these methods (including the one way analysis of variance) alternatives exist that are robust against variance heterogeneity. A modification of (3) has some unattractive properties if the variances and the sample sizes differ greatly. The adaptations for unequal variances of (4) and (5) seem better than (1) for cases with many samples. Test (2) is rather robust in itself if the variances are not too much different. Modifications exist that allow slight unequalities in the sample sizes.

1. INTRODUCTION

In 1981 Werter and the author published a study on tests for the equality of several means when the population variances are unequal. The problem can be stated as follows:

$$H_0: \mu_1 = \dots = \mu_k$$

$$\underline{x}_{ij} \sim N(\mu_i, \sigma_i^2) \text{ for } i = 1, \dots, k \\ j = 1, \dots, n_i.$$

The conclusion of this study was that the second order method of James (1951) gives the user better control over the size than some other tests [Welch (1951), Brown and Forsythe (1974)], so it is to be preferred since none of the tests in the study was uniformly most powerful.

The test statistic t is defined as:

$$t = \sum_{i=1}^k w_i (x_i - \bar{x})^2, \text{ where } w_i = \frac{x_i}{s_i^2}, x_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \bar{x} = \frac{1}{w} \sum_{i=1}^k w_i x_i \text{ and } w = \sum_{i=1}^k w_i.$$

For some chosen size α this test statistic is to be compared with a critical level $h_2(\alpha)$, given by:

$$\begin{aligned} h_2(\alpha) = & \chi^2 + \frac{1}{2}(3\chi_4 + \chi_2) \sum_{i=1}^k \frac{1}{v_i} \left(1 - \frac{w_i}{w}\right)^2 + \left\{ \frac{1}{16} (3\chi_4 + \chi_2)^2 \right. \\ & \left. \left(1 - \frac{k-3}{2}\right) \left(\sum_{i=1}^k \frac{1}{v_i} \left(1 - \frac{w_i}{w}\right)^2 \right)^2 + \frac{1}{2}(3\chi_4 + \chi_2) (8R_{23} - 10R_{22} + 4R_{21} - \right. \\ & 6R_{12}^2 + 8R_{12}R_{11} - 4R_{11}^2) + (2R_{23} - 4R_{22} + 2R_{21} - 2R_{12}^2 + 4R_{12}R_{11} \\ & - 2R_{11}^2)(\chi_2 - 1) + \frac{1}{4}(-R_{12}^2 + 4R_{12}R_{11} - 2R_{12}R_{10} - 4R_{11}^2 + 4R_{11}R_{10} \\ & - R_{10}^2)(3\chi_4 - 2\chi_2 - 1) \left. \right\} + (R_{23} - 3R_{22} + 3R_{21} - R_{20})(5\chi_6 + 2\chi_4 + \\ & \chi_2) + \frac{3}{16}(R_{12}^2 - 4R_{23} + 6R_{22} - 4R_{21} + R_{20})(35\chi_8 + 15\chi_6 + 9\chi_4 + \\ & 5\chi_2) + \frac{1}{16}(-2R_{22} + 4R_{21} + R_{20} + 2R_{12}R_{10} - 4R_{11}R_{10} + R_{10}^2)(9\chi_8 - \\ & 3\chi_6 - 5\chi_4 - \chi_2) + \frac{1}{4}(-R_{22} + R_{11}^2)(27\chi_8 + 3\chi_6 + \chi_4 + \chi_2) + \\ & \left. \frac{1}{4}(R_{23} - R_{12}R_{11})(45\chi_8 + 9\chi_6 + 7\chi_4 + 3\chi_2) \right\} \end{aligned}$$

Here $\chi^2 = \chi^2(\alpha)$ is the percentage point of a χ^2 -distributed variate with $r = k-1$ degrees of freedom, having a tail probability α . The other basic items in the formula are given by:

$$\chi_{2s} = [\chi^2(\alpha)]^s (k-1)(k+1) \dots (k+2s-3) \text{ and } R_{st} = \sum_{i=1}^k \frac{1}{v_i} \left(\frac{w_i}{w}\right)^t, \text{ where } v_i = n_i - 1.$$

$$P = \left(1 - \frac{\alpha}{\binom{k}{2}}\right)^{\binom{k}{2}} < 1 - \binom{k}{2} \frac{\alpha}{\binom{k}{2}} = 1 - \alpha$$

This method is an approximation of order -2 in the v_i to an "ideal" method. Brown and Forsythe (1974) considered the first order method of James (order -1 in the v_i). Their conclusion was that for unequal variances the difference between the nominal size and the actual probability of rejecting the null hypothesis when it is true can be quite impressive. Werter and the author found that this difference almost vanishes if one takes into account the second order terms.

The test as stated gives only the binary result that H_0 is accepted or rejected. If one prefers the tail probability of the test the equation $t = h_2(\alpha)$ has to be solved. Because $h_2(\alpha)$ is monotonous in α this can be done in about ten function evaluations with an acceptable precision of 0.001 in α . In the formula for $h_2(\alpha)$ the terms R_{st} are independent of α , so it is only necessary to recompute the χ_{2s} for every iteration. This version of the test was used on a Burroughs B7700 computer. The average amount of processing time for common cases was about 0.026 sec, so the very complicated formula does not yield an expensive algorithm.

If H_0 is accepted this usually means the end of the analysis. Otherwise it may be of interest to know where the differences lie. For this one has to perform a simultaneous test and it would be nice if this could be done in such a way that α means "The accepted probability of declaring any pair μ_i, μ_j different when in fact they are equal". In the following sections some strategies are worked out for this kind of simultaneous statistical inference.

2. LEAST SIGNIFICANT DIFFERENCE TEST

The method consists of two stages. First $H_0: \mu_1 = \dots = \mu_k$ is to be tested with classical one way analysis of variance. If H_0 is rejected a t-test is to be performed for every pair. This idea originates from Fisher (1935) and it presupposes the variances to be equal.

Fisher suggested using the same α for the t-tests as for the overall analysis of variance. Of course this is not safe in the sense mentioned in the introduction. An alternative to be considered is the Bonferroni idea $\beta = \alpha / \binom{k}{2}$ that is mentioned in Miller (1966). For this the probability that no error is made under H_0 is limited as follows:

$$P = \left(1 - \frac{\alpha}{\binom{k}{2}}\right)^{\binom{k}{2}} < 1 - \binom{k}{2} \frac{\alpha}{\binom{k}{2}} = 1 - \alpha$$

For unequal variances the one way analysis of variance can be replaced by the James second order test. For comparing the pairs there are several possibilities. The situation is called the Behrens-Fisher (1929) problem, and one of the best approximate solutions is Welch's modified t-test (1949). This test has been evaluated by Wang (1971) and he concluded that it gives the user excellent control over the size, whatever the value of the nuisance parameter $\theta = \sigma_i^2 / \sigma_j^2$ may be. The test statistic is

$$t = \frac{x_i - x_j}{\left[\left(\frac{s_i^2}{n_i} \right) + \left(\frac{s_j^2}{n_j} \right) \right]^{\frac{1}{2}}}$$

and the critical level for some chosen size β is given by Students t-distribution with a parameter ν that takes the pattern of the variances into account:

$$\nu_{ij} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right)^2}{\frac{s_i^4}{n_i^2(n_i-1)} + \frac{s_j^4}{n_j^2(n_j-1)}}$$

In most cases ν_{ij} is not an integer, so it has to be replaced by the nearest one. Ury and Wiggins (1971) suggested using this test with the Bonferroni β . The simultaneous confidence intervals for this approach are given by:

$$\mu_i - \mu_j \in \left[x_i - x_j \bar{+} t_{\nu_{ij}}^{\frac{1}{2}\beta} \left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right)^{\frac{1}{2}} \right]$$

There are some alternatives mentioned in the literature. Hochberg (1976) suggested using:

$$\mu_i - \mu_j \in \left[x_i - x_j \bar{+} \gamma_\alpha \left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right)^{\frac{1}{2}} \right]$$

where γ_α is the solution of $\sum_{i=1}^k \sum_{j=i+1}^k P\{(|t_{\nu_{ij}}| > \gamma)\} = \alpha$, in which ν_{ij} comes from Welch's modified t-test.

Tamhane (1977) suggested using Bajernee's (1961) approximate solution of the Behrens-Fisher problem with $\gamma = 1 - (1-\alpha)^{\frac{1}{k-1}}$. This γ has some history and will also be mentioned in the following sections. The confidence intervals become:

$$\mu_i - \mu_j \in [x_i - x_j \mp \{ (t_{v_i}^{\frac{1}{2}\gamma})^2 \frac{s_i^2}{n_i} + (t_{v_j}^{\frac{1}{2}\gamma})^2 \frac{s_j^2}{n_j} \}^{\frac{1}{2}}]$$

Tamhane also suggested using Welch's test with this γ .

In the literature the author has found nine different approximate solutions of the Behrens-Fisher problem and five ideas concerning the size of the separate tests. Every combination can be made, so there is quite a lot of methods one can consider for pairwise comparisons. But to be really safe, in the sense that the probability of declaring any pair different when in fact they are equal should be limited by α , the pairwise size β will become very small. For $k = 15$ and $\alpha = 0.05$ the Bonferroni approach will yield $\beta = 0.00048$, so it becomes almost impossible to reject any pairwise comparison.

Another disadvantage of this approach is the fact that the results have to be represented by a matrix containing symbols for acceptance and rejection. Working at a terminal, as is usually done in applied statistics nowadays, one has to swallow an enormous lot of information in one glance if k exceeds the region of very small values. The next sections will suggest approaches that are better in this respect.

3. MULTIPLE RANGE TESTS

In this section a strategy will be pointed out that was originated by Newman (1939), Duncan (1951) and Keuls (1952). At first it will be necessary for the sample sizes to be equal ($n_i = n$ for $i = 1, \dots, k$). Also variance heterogeneity will not be allowed. Later on these limitations will be dropped.

Let $x_{(1)}, \dots, x_{(k)}$ be the sample means, sorted in non-decreasing order. The first hypothesis of interest is $H_0: \mu_1 = \dots = \mu_k$, where the μ_i 's are renumbered so that their ordering becomes the same as the sample means which are their estimates.

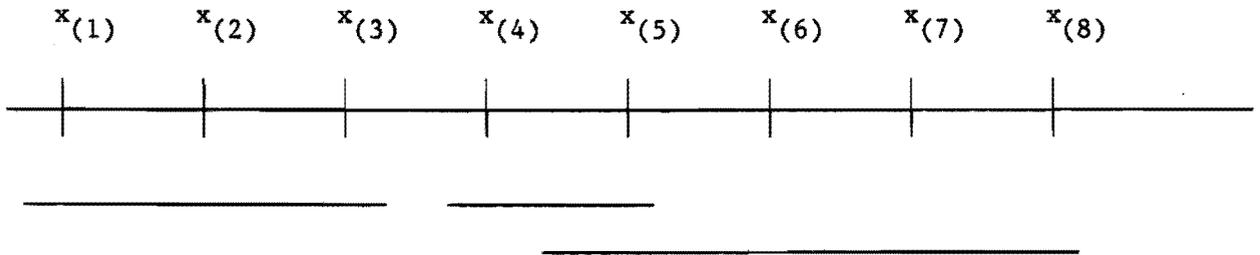
Then H_0 can be tested with:

$$\mu_1 - \mu_k \in [x_1 - x_k \mp q_{k,v}^\alpha \frac{s}{n^{\frac{1}{2}}}]$$

where q is the studentized range distribution, $v = k(n-1)$ and the residual variance is estimated by:

$$s^2 = \frac{1}{v} \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - x_i)^2.$$

If H_0 is rejected, the next stage is to test $\mu_1 = \dots = \mu_{k-1}$ and $\mu_2 = \dots = \mu_k$. Proceeding like this until every hypothesis is accepted will yield a result that can be represented as follows:



The interpretation of this figure is that $\mu_i = \mu_j$ has to be rejected if there is no unbroken line that underscores $x_{(i)}$ and $x_{(j)}$. For instance:

- $\mu_4 = \mu_5$: accepted
- $\mu_5 = \mu_6$: accepted
- $\mu_4 = \mu_6$: rejected.

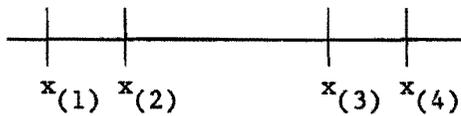
If a candidate for the splitting process contains p means then $q_{p,v}^\alpha$ is to be used instead of $q_{k,v}^\alpha$. Newman and Keuls suggested $\alpha_p = \alpha$ and Duncan preferred $\alpha_p = 1 - (1-\alpha)^{p-1}$.

Now the equality of the sample sizes will be dropped, but for the moment the variances will still have to be equal. Miller (1966) suggested using the median of n_1, \dots, n_k . Winer (1962) considered the harmonic mean $H (\frac{1}{H} = \frac{1}{k} \sum_{i=1}^k \frac{1}{n_i})$.

Kramer (1956) modified the formula of the test to this situation:

$$\mu_i - \mu_j \in [x_i - x_j \pm q_{p, \nu}^{\alpha} s \{ \frac{1}{2} (\frac{1}{n_i} + \frac{1}{n_j}) \}^{\frac{1}{2}}], \text{ where } \nu = N - k \text{ and } N = \sum_{i=1}^k n_i$$

Only in Kramer's case does the studentized range distribution hold. For Miller and Winer the approximation will be reasonable if the sample sizes are not too different. Kramer's test contains a trap that can be shown in the following figure:



Suppose n_1 and n_4 are much smaller than n_2 and n_3 . Then $\mu_1 = \dots = \mu_4$ can be accepted while μ_2 and μ_3 are significantly different. But the strategy will make sure that this difference will never be found.

From here on the variances will be allowed to be unequal. For equal sample sizes Ramseyer and Tcheng (1973) found that the studentized range statistic is remarkably robust against variance heterogeneity. So for almost equal sample sizes it seems reasonable to use the Winer or Miller approach and ignore the differences in the variances.

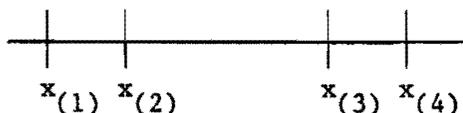
Unfortunately, the robustness of Kramer's test is rather poor [Games and Howell (1976)], so if the sample sizes differ greatly one might be tempted to consider:

$$\mu_i - \mu_j \in [x_i - x_j \pm q_{p, \nu_{ij}}^{\alpha} \{ \frac{1}{2} (\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}) \}^{\frac{1}{2}}]$$

where only the variances of the extreme samples are taken into account. This idea was mentioned by Games and Howell (1976) with Welch's ν_{ij} . The studentized range distribution does not hold for these separately estimated variances, but the approximation seems reasonable though a bit conservative.

The context in which Games and Howell suggested using this method was one of pairwise comparisons with other parameters for q . But it looks like a good start for the construction of a "Generalized Multiple Range test".

This test, however attractive it may seem, still contains the trap that was already mentioned for Kramer's method. But there is more:



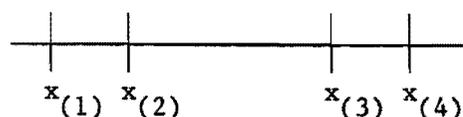
Suppose s_2^2 and s_3^2 are (much) smaller than s_1^2 and s_4^2 . Then a significant difference between μ_2 and μ_3 can easily be ignored.

The author has not found in the literature other approaches to variance heterogeneity within the strategy of multiple range tests. Some other α_p 's have been suggested, but since the choice of α_p has almost nothing to do with robustness against variance heterogeneity, their merits will not be discussed in this paper.

The representation of the results with underscoring lines seems very attractive since this simple figure contains a lot of information, and also the artificial consistency that comes from the ordered means has some appeal. However the whole idea of a Generalized Multiple Range test seems wrong. One simply cannot afford to take only the extreme means into account if the sample sizes and the variances differ greatly.

4. MULTIPLE F-TEST

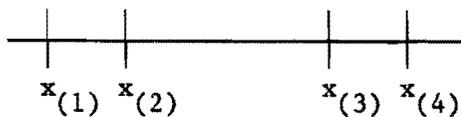
This test was proposed by Duncan (1951). In the original version the population variances must be equal. The procedure is the same as for the Multiple Range test, only the q -statistic is replaced by an F , so that the first stage becomes classical one way analysis of variance. At first Duncan proposed using $\alpha_p = 1 - (1-\alpha)^{p-1}$, but later he found $\alpha_p = 1 - (1-\alpha)^{(p-1)/(k-1)}$ more suitable [Duncan (1955)]. The nature of the F -test allows unequal sample sizes. This seems to make this approach more attractive than the Multiple Range test, but there is a problem:



Suppose $\mu_1 = \dots = \mu_4$ is rejected. The next two hypotheses to be tested are $\mu_1 = \dots = \mu_3$ and $\mu_2 = \dots = \mu_4$. So μ_1 and μ_4 will always be called different. But if n_1 and n_4 are much smaller than n_2 and n_3 it is possible that a pairwise test for μ_1 and μ_4 would not yield any significance. Duncan (1952) saw this problem and suggested using a t-test for the pairs that seemed significant as a result of the Multiple F-test. This approach he called the Multiple Comparisons test. Nowadays this term has a more general meaning and it seems to cover every classifying procedure one might consider after rejecting $\mu_1 = \dots = \mu_k$.

Now the equality of the variances will be dropped. It is well known that the F-test is not robust against variance heterogeneity [Brown and Forsythe (1974), Ekbohm (1976)]. So it seems reasonable to use the non-iterative version of the second order method of James, thus making a "Multiple James test". One could use Duncan's α_p , but the author prefers $\alpha_p = 1 - (1-\alpha)^{p/k}$ [Ryan (1960)] as a consequence of some arguments pointed out by Einot and Gabriel (1975). This α_p was mentioned in another context, but the arguments are not much shaken by the inequality of the variances.

This new test contains the same problem as the Multiple F-test, but that is not all:



μ_1 and μ_4 will always be called different if $\mu_1 = \dots = \mu_4$ is rejected. Now suppose that s_2^2 and s_3^2 are much smaller than s_1^2 and s_4^2 . Then the difference between μ_1 and μ_4 may not be significant in a pairwise comparison. Here the structural difference between this test and the approach mentioned in the previous section comes into the picture: If extreme means coincide with big variances and small samples, then the Generalized Multiple Range test can ignore important differences, while the Multiple James test can wrongly declare means to be different.

One can of course apply Welch's test for the Behrens-Fisher problem to the pairs that seem significant as a consequence of the Multiple James test. This combination should be called the "Generalized Multiple Comparisons test". A lot of extra work may be asked for, so it is of interest to know if this extension can have any serious influence on the conclusions.

Werter and the author have examined this by adding another member to the family: the "Leaving One Out test". This is a Multiple James test in which after rejection of $\mu_1 = \dots = \mu_k$ not only $\mu_1 = \dots = \mu_{k-1}$ and $\mu_2 = \dots = \mu_k$ are considered but all the subsets of μ_1, \dots, μ_k where one μ_i is left out. The same α_p is used and the acceptance of a hypothesis means that the splitting process for this subset stops. The Leaving One Out strategy is not limited to $\mu_1 = \dots = \mu_k$ but is applied to every subset that becomes a candidate. This approach will avoid the classical trap of the Multiple F-test and also the specific problem that comes from variance heterogeneity.

The Multiple James test and the Leaving One Out test were applied to 7 case studies, containing 277 pairs. Only 2 different pairwise conclusions were reached, where the Leaving One Out test did not confirm the significance found by the Multiple James test. But since the Multiple Comparisons test is considered a useful extension of the Multiple F-test, this may not be representative.

The Leaving One Out test can be very expensive. In the worst case situation where all the means are isolated the number of tests will be $2^k - (k+1)$ instead of only $\frac{1}{2}k(k-1)$ for the Multiple James test and any member of the Least Significant Difference family. For $k = 15$ this means 32752 tests instead of only 105.

For values of k that make the Least Significance Difference approach unattractive, the Multiple James test is recommended with Ryan's α_p . A terminal oriented computer program such as BMDP should not only give the final result but also the mean, variance and number of observations for every sample. An interesting pairwise significance can be verified by Welch's test for the Behrens-Fisher problem. This should be considered if the sample variances involved are relatively big or if the samples contain only a few observations.

5. FINAL REMARK

This small study on robustness of multiple comparisons against variance heterogeneity only just touches some of the major problems. They are dealt with separately in a simplified example of four samples. In reality one has to deal with them simultaneously which makes the problems much more difficult. Also there are some well known disturbing effects that are not mentioned in this paper.

6. ACKNOWLEDGEMENTS

Prof.dr. R. Doornbos has been willing to discuss this study in two stages of its development. For his comments I am very grateful.

I wish to thank Paul Werter for his simulation study on the robustness of the q-statistic as used in the Generalized Multiple Range test. And also for comparing the behaviour of the robust tests of the last section in seven case studies.

Further I wish to express my gratitude for the helpfulness of Marjan van Rooij who typed this manuscript in a hurry.

7. REFERENCES

- [1] Dijkstra, Jan B. and Paul S.P.J. Werter
Testing the equality of several means when the population variances are unequal
Communication in Statistics (1981) B 10, 6
- [2] Brown, M.B. and A.B. Forsythe
The small sample behaviour of some statistics which test the equality of several means
Technometrics 16 (1974), 129-132
- [3] Welch, B.L.
The comparison of several mean values: an alternative approach
Biometrika 38 (1951), 330-336
- [4] James, G.S.
The comparison of several groups of observations when the ratios of the population variances are unknown
Biometrika 38 (1951), 324-329
- [5] Fisher, R.A.
The Design of Experiments
Oliver & Boyd (1935) Edinburgh and London
- [6] Miller, R.G.
Simultaneous Statistical Inference
McGraw Hill Book Company, New York (1966)
- [7] Behrens, W.V.
Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen
Landwirtschaftliche Jahrbücher 68 (1929), 807-837

- [8] Welch, B.L.
Further note on Mrs. Aspin's tables and on certain approximation to the tabled function
Biometrika 36 (1949), 293 - 296
- [9] Wang, Y.Y.
Probabilities of type I errors of the Welch tests for the Behrens-Fisher problem
Journal of the American Statistical Association 66 (1971)
- [10] Ury, H.K. and A.D. Wiggins
Large sample and other multiple comparisons among means
British Journal of Mathematical and Statistical Psychology 24 (1971), 174-194
- [11] Hochberg, Y.
A modification of the T-method of multiple comparisons for a one-way layout with unequal variances
Journal of the American Statistical Association 71 (1976), 200-203
- [12] Tamhane, A.C.
Multiple Comparisons in model-1 one way ANOVA with unequal variances
Communications in Statistics A6(1) (1977), 15-32
- [13] Banerjee, S.K.
On confidence intervals for two-means problem based on separate estimates of variances and tabulated values of t-variable
Sankhya, A23 (1961)
- [14] Newman, D.
The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation
Biometrika 31 (1939), 20-30
- [15] Keuls, M.
The use of the "studentized range" in connection with an analysis of variance
Euphytica 1 (1952), 112-122
- [16] Winer, B.J.
Statistical principles in experimental design
New York, McGraw-Hill (1962)
- [17] Kramer, C.Y.
Extension of multiple range tests to group means with unequal numbers of replications
Biometrics 12 (1956), 307-310

- [18] Ramseyer G.C. and T. Tcheng
The robustness of the studentized range statistic to violations of the normality and homogeneity of variance assumptions
American Educational Research Journal 10 (1973)
- [19] Games, P.A. and J.F. Howell
Pairwise multiple comparison procedures with unequal N's and/or Variances: a Monte Carlo Study
Journal of Educational Statistics 1 (1976), 113-125
- [20] Duncan, D.B.
A significance test for differences between ranked treatments in an analysis of variances
Virginia Journal of Science 2 (1951)
- [21] Duncan, D.B.
Multiple range and Multiple F-tests
Biometrics 11 (1955), 1-42
- [22] Duncan, D.B.
On the properties of the multiple comparisons test
Virginia Journal of Science 3 (1952)
- [23] Ekbohm, G.
On testing the equality of several means with small samples
The Agricultural College of Sweden, Uppsala (1976)
- [24] Ryan, T.A.
Significance Tests for multiple comparison of proportions, variances and other statistics
Psychological Bulletin 57 (1960), 318-328
- [25] Einot, I. and K.R. Gabriel
A study of the powers of several methods of multiple comparisons
Journal of the American Statistical Association 70 (1975), 574-583