Comparison of nonparametric analysis of variance methods a Monte Carlo study
Part A: Between subjects designs - A Vote for van der Waerden

Version 5
completely revised and extended
(13.7.2017)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Contact: Luepsen@Uni-Koeln.de

Universität zu Köln

Comparison of nonparametric analysis of variance methods - a Vote for van der Waerden

Abstract

For two-way layouts in a between subjects anova design the parametric F-test is compared with seven nonparametric methods: rank transform (RT), inverse normal transform (INT), aligned rank transform (ART), a combination of ART and INT, Puri & Sen's L statistic, van der Waerden and Akritas & Brunners ATS. The type I error rates and the power are computed for 16 normal and nonnormal distributions, with and without homogeneity of variances, for balanced and unbalanced designs as well as for several models including the null and the full model. The aim of this study is to identify a method that is applicable without too much testing all the attributes of the plot. The van der Waerden-test shows the overall best performance though there are some situations in which it is disappointing. The Puri & Sen- and the ATS-tests show generally a very low power. These two as well as the other methods cannot keep the type I error rate under control in too many situations. Especially in the case of lognormal distributions the use of any of the rank based procedures can be dangerous for cell sizes above 10. As already shown by many other authors, nonnormal distributions do not violate the parametric F-test, but unequal variances do. And heterogeneity of variances leads to an inflated error rate more or less also for the nonparametric methods. Finally it should be noted that some procedures show rising error rates with increasing cell sizes, the ART, especially for discrete variables, as well as the RT, Puri & Sen and the ATS in the cases of heteroscedasticity.

Keywords: nonparametric anova, rank transform, Puri & Sen, ATS, Waerden, simulation

1. Introduction

The analysis of variance (anova) is one of the most important and frequently used methods of applied statistics. In general it is used in its parametric version often without checking the assumptions. These are normality of the residuals, homogeneity of the variances - there are several different assumptions depending on the design - and the independence of the observations. Most people trust in the robustness of the parametric tests. „*A test is called robust when its significance level (Type I error probability) and power (one minus Type-II probability) are insensitive to departures from the assumptions on which it is derives.*" (See Ito, 1980). Good reviews of the assumptions and the robustness can be found at Field (2009) and Ito (1980), more detailed descriptions at Fan (2006), Wilcox (2005), Osborne (2008), Lindman (1974) as well as Glass et al. (1972). They state that first the F-test is remarkable insensitive to general nonnormality, and second the F-test can be used with confidence generally when variances are equal, as well as in cases of variance heterogeneity at least in cases with equal sample sizes. However Box (1954), Glass et al. (1972) and Dijkstra (1987) have shown that even in balanced designs unequal variances may lead to an increased type I error rate. It remains to mention one severe problem of the F-test for the case of unequal n_i : it tends to be conservative if cells with larger n_i have also larger variances (positive pairing) and that it reacts liberal if cells with larger n_i have the smaller variances (negative pairing), as Feir & Toothaker (1974) and Weihua Fan (2006), to name a few, reported. Nevertheless other methods may exist which are superior in these cases even when the F-test may be applicable. Furthermore dependent variables with an ordinal scale normally require adequate methods.

The knowledge of nonparametric methods for the anova is not wide spread though in recent

years quite a number of publications on this topic appeared. Salazar-Alvarez et al. (2014) gave a review of the most recognized methods. Another easy to read review is one by Erceg-Hurn and Mirosevich (2008). As Sawilowsky (1990) pointed out, it is often objected that nonparametric methods do not exhaust all the information in the data. This is not true. Sawilowsky (1990) also showed that most well-known nonparametric procedures, especially those considered here, have a power comparable to their parametric counterparts, and often a higher power when assumptions for the parametric tests are not met.

On the other side are nonparametric methods not always acceptable substitutes for parametric methods such as the F-test in research studies when parametric assumptions are not satisfied. „*It came to be widely believed that nonparametric methods always protect the desired significance level of statistical tests, even under extreme violation of those assumptions*“ (see Zimmerman, 1998). Especially in the context of anova with the assumptions of normality and variance homogeneity. And there exist a number of studies showing that nonparametric procedures cannot handle skewed distributions in the case of heteroscedasticity (see e.g. G. Vallejo et al., 2010, Keselman et al., 1995 and Tomarken & Serlin, 1986).

A barrier for the use of nonparametric anova is apparently the lack of procedures in the statistical packages, e.g. SAS and SPSS though a number of SAS macros meanwhile exist. For R and S-Plus packages with corresponding algorithms have been supplied during the last years. But as is shown by Luepsen (2015) a number of the nonparametric anova methods can be applied by using the parametric standard anova procedures together with a little bit of programming, for instance to do some variable transformations. Such algorithms stay in the foreground.

The aim of this study is to identify situations, e.g. designs or underlying distributions, in which one method is superior compared to others. For, many appliers of the anova know only little of their data, the shape of the distribution, the homogeneity of the variances or expected size of the effects. So, overall good performing methods are looked for. But attention is also laid upon comparisons with the F-test. As usual this is achieved by examining the type I error rates at the 5 and 1 percent level as well as the power of the tests at different levels of effect or sample size. Here the focus is laid not only upon the tests for the interaction effects but also on the main effects as the properties of the tests have not been studied exhaustively in factorial designs. Additionally the behavior of the type I error rates is examined for increasing cell sizes up to 50, because first, as a consequence of the central limit theorem, some error rates should decrease for larger n_i, and second most nonparametric tests are asymptotic. The present study is concerned only with between subjects designs.

2. Methods to be compared

It follows a brief description of the methods compared in this paper, none of them considering heterogeneous variances. More information, especially how to use them in R or SPSS can be found in Luepsen (2015).

The parametric F-test

The 2-factorial anova model for a dependent variable y with N observations shall be denoted by

$$y_{ijk} = \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$$

with fixed effects α_i (factor A, i=1,..,I), β_j (factor B, j=1,..,J), $\alpha\beta_{ij}$ (interaction AB), error e_{ijk} (k=1,..,n_{ij}), cell counts n_{ij} and $N=\sum n_{ij}$. The parameters α_i , β_j and $\alpha\beta_{ij}$ with the restrictions

$\sum \alpha_i = 0$, $\sum \beta_j = 0$, $\sum \alpha\beta_{ij} = 0$ can be estimated by means of a linear model $\mathbf{y'} = \mathbf{X}\,\mathbf{p'} + \mathbf{e'}$ using the least squares method, where **y** are the values of the dependent variable, **p** is the vector of the parameters, **X** a suitable design matrix and **e** the random variable of the errors. If the contrasts for the tests of the hypotheses H_A (α_i=0), H_B (β_i=0) and H_{AB} ($\alpha\beta_{ij}$=0) are orthogonal the resulting sum of squares SS_A, SS_B, SS_{AB} of the parameters are also orthogonal and commonly called type III SSq. They are tested by means of the F-distribution. In case of equal sample sizes the sum of squares as well as the mean squares can be easily computed as

$$SS_A = \frac{N}{I}\sum(\bar{y}_{i\cdot\cdot} - \bar{y})^2 \qquad SS_B = \frac{N}{J}\sum(\bar{y}_{\cdot j\cdot} - \bar{y})^2 \qquad SS_{AB} = \frac{N}{IJ}\sum\sum(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y})^2$$

$$MS_A = SS_A/(I-1) \qquad MS_B = SS_B/(J-1) \qquad MS_{AB} = SS_{AB}/((I-1)(J-1))$$

$$MS_{error} = \sum\sum\sum(y_{ijk} - \bar{y}_{ij\cdot})^2/(N-IJ)$$

and the F-ratios as

$$F_A = MS_A/MS_{error} \qquad F_B = MS_B/MS_{error} \qquad F_{AB} = MS_{AB}/MS_{error}$$

where $\bar{y}_{i\cdot\cdot}$, $\bar{y}_{\cdot j\cdot}$ are the level means of factor A and B, $\bar{y}_{ij\cdot}$ are the cell means and $\bar{y}$ is the grand mean (see e.g. Winer, 1991).

RT (rank transform)

The rank transform method (RT) is just transforming the dependent variable *y* into ranks R(*y*) before applying the parametric F-test, as described above, to them. This method had been proposed by Conover & Iman (1981). Blair et al. (1987), Toothaker & Newman (1994) as well as Beasley & Zumbo (2009), to name only a few, found out that the type I error rate of the interaction can reach beyond the nominal level if there are significant main effects because the effects are confounded. On the other hand the RT lets sometimes vanish an interaction effect, as Salter & Fawcett (1993) had shown in a simple example. The reason: *„additivity in the raw data does not imply additivity of the ranks, nor does additivity of the ranks imply additivity in the raw data“,* as Hora & Conover (1984) pointed out. At least Hora & Conover (1984) proved that the tests of the main effects are correct. A good review of articles concerning the problems of the RT can be found in the study by Toothaker & Newman (1994).

INT (inverse normal transform)

The inverse normal transform method (INT) consists of first transforming *y* into ranks R(*y*) (as in the RT method), then computing their normal scores and finally applying the parametric F-test to them. The normal scores are defined as

$$\Phi^{-1}(R(y)/(N+1))$$

where R(*y*) are the ranks of *y* and *N* is the number of observations. It should be noted that there exist several versions of the normal scores (see Beasley, Erickson & Allison, 2009, for details). This results in an improvement of the RT procedure as could be shown by Huang (2007) as well as Mansouri and Chang (1995), though Beasley, Erickson & Allison (2009) found out that also the INT procedure results in slightly too high type I error rates if there are other significant main effects. This method should not be confused with the expected normal scores test by Hájek et al. (1999).

ART (aligned rank transform)

In order to avoid an increase of type I error rates for the interaction in case of significant main effects an alignment is proposed: all effects that are not of primary interest are subtracted before

performing an anova. The procedure consists of first computing the residuals, either as differences from the cell means or by means of a regression model, then adding the effect of interest, transforming this sum into ranks and finally performing the parametric F-test to them. For the alignment in the simple 2-factorial model first the error e_{ijk} is computed as the residuals from the parametric anova including all effects as described above. In the next step the means corresponding to the effect being tested (A, B and AB) are added:

$$y_{ijk}^{(A)} = e_{ijk} + a_i \qquad y_{ijk}^{(B)} = e_{ijk} + b_j \qquad y_{ijk}^{(AB)} = e_{ijk} + ab_{ij}$$

where a_i, b_j, ab_{ij} are the means of y corresponding to the effect. Then the aligned variables $y^{(A)}$,$y^{(B)}$, $y^{(AB)}$ are each transformed into ranks R($y^{(A)}$), R($y^{(B)}$), R($y^{(AB)}$). To test one effect the parametric F-test is applied to the corresponding aligned variable where only that effect is examined ignoring the other two.

As the normal theory F-tests are used for testing these rank statistics the question arises if their asymptotic distribution is the same. Salter & Fawcett (1993) showed that at least for the ART these tests are valid.

Yates (2008) and Peterson (2002) among others went a step further and used the median as well as several other robust mean estimates for adjustment in the ART-procedure. Besides this there exist a number of other variants of alignment procedures. for example the M-test by McSweeney (1967), the H-Test by Hettmansperger (1984) and the RO-test by Toothaker & De Newman (1994). But in a comparison by Toothaker & De Newman (1994) the latter three showed a liberal behavior. Because of this and the fact that they are not widespread these procedures had not been taken into consideration for this study.

This procedure can also be applied to the test of main effects - which is done in this study - though this is not necessary as mentioned above.

ART combined with INT (ART+INT)

Mansouri & Chang (1995) suggested to apply the normal scores transformation INT (see above) to the ranks obtained from the ART procedure. They showed that the transformation into normal scores improves the type I error rate, for the RT as well as for the ART procedure, at least in the case of underlying normal distributions. Computationally the steps are nearly the same as for the ART method above, with the difference that the ranked aligned variables R($y^{(A)}$), R($y^{(B)}$), R($y^{(AB)}$) are transformed into normal scores, as described for the INT method, before applying the parametric F-test on them.

Puri & Sen tests (L statistic)

These are generalizations of the well known Kruskal-Wallis H test (for independent samples) and the Friedman test (for dependent samples) by Puri & Sen (1985), often referred as L statistic. A good introduction offer Thomas et al. (1999). The idea dates back to the 60s, when Bennett (1968) and Scheirer, Ray & Hare (1976) as well as later Shirley (1981) generalized the H test for multifactorial designs. It is well-known that the Kruskal-Wallis H test as well as the Friedman test can be performed by a suitable ranking of y (see e.g. Winer,1991), conducting a parametric anova and finally computing χ^2-ratios using the sum of squares. In fact the same applies to the generalized tests. In the simple case of only grouping factors the χ^2-ratios for the tests of A, B and AB are computed as

$$\chi_A^2 = \frac{SS_A}{MS_{total}} \qquad \chi_B^2 = \frac{SS_B}{MS_{total}} \qquad \chi_{AB}^2 = \frac{SS_{AB}}{MS_{total}}$$

Here SS_A, SS_B, SS_{AB} are the sum of squares as outlined before, but computed for R(*y*), the ranks of *y*, and MS_{total} is the total mean square, i.e. the variance of R(*y*). The degrees of freedom are those of the numerator of the corresponding F-test.

The major disadvantage of this method compared with the four ones above is the lack of power for any effect in the case of other nonnull effects in the model. The reason: In the standard anova the denominator of the F-values is the residual mean square which is reduced by the effects of other factors in the model. In contrast the denominator of the χ^2 tests of Puri & Sen's L statistic is the total mean square which increases with effects of the other factors, thus making the ratio of the considered effect and therefore also the χ^2-ratio smaller. A good review of articles concerning this test can be found in the study by Toothaker & De Newman (1994).

van der Waerden

At first the van der Waerden test (van der Waerden,1953) is an alternative to the 1-factorial anova by Kruskal-Wallis. The procedure is based on the INT transformation (see above). But instead of using the F-tests from the parametric anova χ^2-ratios are computed using the sum of squares in the same way as for the Puri & Sen L statistics. Mansouri and Chang (1995) generalized the original van der Waerden test to designs with several grouping factors. Computationally the steps are nearly identical with those above for the L statistic, with the difference that the ranks R(*y*) are transformed into normal scores as described for the INT-method above before computing the sum of squares and χ^2-ratios.

Perhaps it is to mention that Sheskin (2004) reported that this procedure in its 1-factorial version outperforms the classical anova in the case of violations of the assumptions with regard to the power. And Hajek (1969) showed that this test has asymptotically the same efficiency as the F-test. On the other hand the van der Waerden tests suffer from the same lack of power in the case of multifactorial designs as the Puri & Sen L statistic.

Akritas, Arnold and Brunner (ATS)

This is the only procedure considered here that cannot be mapped to the parametric anova. Based on the relative effect (see Brunner & Munzel (2002)) the authors developed two tests to compare samples by means of comparing the relative effects: the approximately F distributed ATS (anova type statistic) and the asymptotically χ^2 distributed WTS (Wald type statistic). The ATS has preferable attributes e.g. more power (see Brunner & Munzel (2002) as well as Shah & Madden (2004)). The relative effect of a random variable X_1 to a second one X_2 is defined as $p^+ = P(X_1 \leq X_2)$, i.e. the probability that X_1 has smaller values than X_2 . As the definition of relative effects is based only on an ordinal scale of *y* this method is suitable also for variables of ordinal or dichotomous scale. The rather complicated procedure involves a lot of matrix algebra and is described in the appendix.

Methods dropped from this study

In the preceding sections a couple of methods had been mentioned that had not been considered in this study, mainly because of a violation of the type I error rates. For the same the following tests were dropped from this study. For detailed error rates see tables in appendix A 1.6 and A 1.7, for the power of the test by Gao & Alvo see A 3.15.

- The Wilcoxon analysis (WA) that had been proposed by Hettmansperger and McKean (2011) and for which there exists also the R package `Rfit` (see Terpstra & McKean, 2005). WA is primarily a nonparametric regression method. It is based on ranking the residuals and

minimizing the impact that extreme values of *y* have on the regression line. Trivially this method can be also used as a nonparametric anova.

- Gao & Alvo (2005) proposed a nonparametric test for the interaction in 2-way layouts for which also a function exists in the R package `StatMethRank` (see Li Qinglong (2015)). This method is fairly liberal with superior power rates especially for small sample sizes at the cost of high type I error rates near 9 percent (at a nominal level of 5 percent) in the case of the null model.

Methodological remarks

Concerning the ATS it should be noted that the name is not unique. It rather denotes an approximately F distributed statistic (anova type), normally derived from a χ^2-statistic, often called WTS. In contrary to the WTS the ATS accounts for the sample sizes that makes it attractive for small cell counts. For instance there exist two variations of this method by Brunner, Dette and Munk (1997), also called BDM-tests, which allow heterogeneous variances: a parametric and a nonparametric one. Both are also available as ATS and WTS tests and for which meanwhile also exist R packages: `GFD` (see Friedrich et al., 2017) and `asbio`. Richter & Payton (2003a) combined the above mentioned ATS with the ART procedure in the way that the BDM-test is applied to the aligned data. In a simulation they showed that this method is better in controlling the type I error rate. These tests are not part of this study though preliminary studies showed that the parametric version of the BDM-test (R function GFD) has not the deficiency of exceeding error rates for rising cell counts n_{ij} , but at the cost of an even lower power than that of the ATS used in this study.

An interesting fact is that Brunner & Puri (2002) developed an approach to test nonparametric hypotheses in a factorial design based on score functions for data that may come from continuous as well as from discrete ordinal distributions. Many well known rank statistics are special cases of this procedure, among others the above mentioned RT, INT, Puri & Sen, v.d.Waerden and ATS, e.g. by choosing Wilcoxon or normal scores as score function. In general the statistics follow asymptotically a χ^2-distribution. However, the speed of approximation to the χ^2-distribution is rather slow, especially if the number of factor levels is large. Therefore Brunner & Puri modified the statistic for small samples so that it could be approximated by an F distribution. Additionally they proposed F tests based on a modified Box approximation (Box, 1954) to allow for heteroscedastic samples. These result in anova type statistics (ATS). These are the versions of the statistics described in the previous section. By the way, Danbaba (2012) compared the two versions of the tests for several score functions, e.g. normal scores. He found too large type I error rates of the χ^2-tests for main effects for small samples ≤ 10 and for the interaction effect even for larger samples ≤ 40 while the F tests kept the error rate generally under control.

Finally some remarks on the ART outlined above. This procedure dates back to Hodges & Lehmann (1962) who developed an alignment similar as outlined above resulting in the Aligned Rank Test (AR), a statistic that follows asymptotically a χ^2-distribution. It is described in detail by Hájek, Šidák and Sen (1999). Mehra & Sarangi (1967) showed that the AR has an asymptotic relative efficiency compared to the normal anova F test of almost 1, and larger than 1 compared to the RT which lets the statistic look attractive. Here also the AR can be transformed for small samples into a statistic that can be approximated by the F distribution resulting in the aligned rank transform (ART) method. This one had been made popular by Higgins & Tashtoush (1994) who extended it to factorial designs.

3. Literature Review

One should be aware that hypotheses of the nonparametric anova tests are not always the same as those of the parametric anova, e.g. the nonparametric hypothesis of no interaction does not imply no interaction in the linear model. Therefore the results achieved for those and mentioned below need not be always applicable one-to-one to the parametric hypothesis.

The ART procedure seems to be the most popular nonparametric anova method judging from the number of publications. But in most papers its behavior is examined only for the comparison of normal and nonnormal distributions in relation to the parametric F-test and the RT method. Some of their results shall be reported first.

Generally the ART-method is to be preferred to the aligned rank (AR) because of its better control of the type I error rate and the larger power (Mansouri 1999a and 1999b). The ART-technique has been estimated rather good in general by Lei et al. (2004), Wobbrock et al. (2011) and Mansouri et al. (2004) to name only a few. Higgins & Tashtoush (1994), Mansouri (1999a) as well as Salter & Fawcett (1993) showed that the ART procedure is valid concerning the type I error rate and preferable to the F-test in cases of outliers or heavily tailed distributions, as in these situations the ART has a larger power than the F-test. Mansouri et al. (2004) studied the influence of noncontinuous distributions and showed the ART to be robust. Richter & Payton (1999) compared the ART with the F-test and with a rank test using the exact permutation distribution, but only to check the influence of violation of normal assumption. For nonnormal distributions the ART is superior especially using the exact probabilities.

There are only few authors who investigated also its behavior in heteroscedastic conditions. Among those are Leys & Schumann (2010) and Carletti & Claustriaux (2005). The first analyzed 2*2 designs for various distributions with and without homogeneity of variances. They found that in the case of heteroscedasticity the ART has even more inflated type I errors than the F-test and that concerning the power only for the main effects the ART can compete with the classical tests. Carletti & Claustriaux (2005) who used a 2*4 design with a relation of 4 and 8 for the ratio of the largest to the smallest variance came to the same results. In addition the type I error increases with larger cell counts. But they proposed an amelioration of the ART technique: to transform the ranks obtained from the ART into normal scores (see 2.4). This method leads to a reduction of the type I error rate, especially in the case of unequal variances.

The use of normal scores instead of ranks had been suggested many years ago by Mansouri & Chang (1995). They showed not only that the ART performs better than the F-test concerning the power in various situations with skewed and tailed distributions but also that the transformation into normal scores improves the type I error rate, for the RT as well as for the ART procedure (resulting in INT and ART+INT), at least in the case of underlying normal distributions. They stated also that none of these is generally superior to the others in any situation. Lachenbruch & Clements (1991) prefer the normal scores to the F-test because of their power in the cases of nonnormality and heteroscedasticity. Concerning the INT-method a long critical disquisition on it by Beasley et al. (2009) exists with a large list of studies dealing with this procedure. They conclude that there are some situations where the INT performs perfectly, e.g. in the case of extreme nonnormal distributions, but there is no general advice for it because of other deficiencies.

Patrick (2007) compared the parametric F-test, the Kruskal-Wallis H-test and the F-test based on normal scores for the 1-factorial design. He found that the normal scores perform the best concerning the type I error rate in the case of heteroscedasticity, but have the lowest power in

that case. By the way he offers also an extensive list of references. A similar study regarding these tests for the case of unequal variances, together with the anovas for heterogeneous variances by Welch and by Brown & Forsythe, comes from Tomarken & Serlin (1986). They reported that the type I error rate as well as the power are nearly the same for the H-test and the INT-procedure. Beside these there exist quite a number of papers dealing with the situation of unequal variances, but unfortunately only for the case of a 1-factorial design, mainly because of lack of tests for factorial designs, as already mentioned above. One of these by Richter & Payton (2003a) who compare the F-test with the ATS and find that the ATS is conservative but always keeps the α-level, by Lix et al. (1996) who compare the same procedures as Tomarken & Serlin did, and by Konar et al. (2015) who compare the one-way anova F-test with Welch's anova, Kruskal Wallis test, Alexander-Govern test, James-Second order test, Brown-Forsythe test, Welch's heteroscedastic F-test with trimmed means and Winsorized variances and Mood's Median test.

Among the first who compared a nonparametric anova with the F-test were Feir & Toothaker (1974) who studied the type I error as well as the power of the Kruskal-Wallis H-test under a large number of different conditions. As the K-W test is a special case of the Puri & Sen method their results are here also of interest: In general the K-W test keeps the α level as good as the F-test, in some situations, e.g. negatively correlating n_i and s_i , even better, but at the cost of its power. The power of the K-W test often depends on the specific mean differences, e.g. if all means differ from each other or if only one mean differs from the rest. Nonnormality has in general little impact on the differences between the two tests, though for an underlying (skewed and tailed) exponential distribution the power of the K-W test is higher. Another interesting paper is the already above mentioned one by Toothaker and De Newman (1994). They compared the F-test with the Puri & Sen test, the RT and the ART method. The Puri & Sen test controls always the type I error but is rather conservative, if there are also other nonnull effects. On the other hand, as the effects are confounded when using the RT method, Toothaker and De Newman propagate the ART procedure for which they report several variations. But all these are too liberal in quite a number of situations. Therefore the authors conclude that there is no general guideline for the choice of the method.

Only a few publications deal with the properties of the ATS method. Hahn et al. (2014) investigated this one together with several permutation tests under different situations and confirmed that the ATS always keeps the α level and that it reacts generally rather conservative, especially for smaller sample sizes (see also Richter & Payton, 2003b). Another study by Kaptein et al. (2010) showed, unfortunately only for a 2*2-design, the power of the ATS being superior to the F-test in the case of Likert scales.

Comparisons of the Puri & Sen L method, the van der Waerden tests or Akritis and Brunner's ATS with other nonparametric methods are very rare. At this point two studies have to be mentioned: First Danbaba (2009) compared for a simple 3*3 two-way design 25 rank tests with the parametric F-test. He considered 4 distributions but unfortunately not the case of heterogeneous variances. His conclusion: among others the RT, INT, Puri & Sen and ATS fulfill the robustness criterion and show a power superior to the F-test (except for the exponential distribution) whereas the ART fails. Secondly Dijkstra (1987) analyzed a large number of solutions for the 1-factorial anova in non-standard situations and remarked the general good performance of the van der Waerden test in settings with unequal variances and nonnormal distributions. So this present study tries to fill some of the gaps.

4. Methodology of the study

General design

This is a pure Monte Carlo study. That means a couple of designs and theoretical distributions had been chosen from which a large number of samples had been drawn by means of a random number generator. These samples had been analyzed using the various anova methods. Concerning the number of different situations, e.g. distributions, equal/unequal variances, equal/ unequal cell counts, effect sizes, relations of means, variances and cell counts, one had to restrict to a minimum, as the number of resulting combinations produce an unmanageable amount of information. Therefore not all influencing factors could be varied. E.g. Feir & Toothaker (1974) had chosen for their study on the Kruskal-Wallis test: two distributions, six different cell counts, two effect sizes, four different relations for the variances and five significance levels. Concerning the results nearly every different situation, i.e. every combination of the settings, brought a slightly different outcome. This is not really helpful from a practical point of view. But on the other side one has to be aware that the present conclusions are to be generalized only with caution. For, as Feir & Toothaker among others had shown, the results are dependent e.g. on the relations between the cell means (order and size), between the cell variances and on the relation between the cell means and cell variances. Own preliminary tests confirmed the influence of the design (number of cells and cell sizes), the pattern of effects as well as size and pattern of the variances on the type I error rates as well as on the power rates.

The current study with two grouping (between subjects) factors A and B examines:

- two layouts:
  - a 2*4 balanced design with 10 observations per cell (total N=80) and
  - a 4*5 unbalanced design with an unequal number of observations n_{ij} per cell (total N=100) and a ratio $\max(n_{ij})/\min(n_{ij})$ of 4,

  which differ not only regarding the cell counts but also the number of cells, though the degrees of freedom of the error term in both designs are nearly equal,
- various underlying distributions (see details below),
- several models for the main and interaction effects.

(In the following sections the terms *unbalanced design* and *unequal cell counts* will be used both for the second design, being aware that they have different definitions. But the special case of a balanced design with unequal cell counts will not be treated in this study.)

The following distributions had been chosen, where the numbers refer also to the corresponding sections in appendix A and where S denotes the skewness:

1. Normal distribution (N(0,1)) with equal variances.
2. N(0,1) with unequal variances with a ratio $\max(s_j^2)/\min(s_j^2)$ of 4 on factor B (correlation n_j with s_j^2 r=0.12).
3. N(0,1) with unequal variances with a ratio $\max(s_{ij}^2)/\min(s_{ij}^2)$ of 4 on both factors (correlation n_{ij} with s_{ij}^2 r=0.04).
4. Right skewed (S~0.8) with equal variances (transformation 1/(0.5+x) with (0,1) uniform x).
5. Exponential distribution (parameter λ=0.4) with μ=2.5 which is extremely skewed (S=2).
6. Exponential distribution (parameter λ=0.4) with μ=2.5 rounded to integer values 1,2,..
7. Lognormal distribution (parameters μ=0 and σ=0.25) which is slightly skewed (S=0.778).

8. Uniform distribution in the interval (0,5).
9. Uniform distribution with integer values 1,2,...,5.
   (First uniformly distributed values in the interval (0,5) are generated, then effects are added and finally rounded up to integers.)
10. Left and right skewed (transformation $\log_2(1+x)$ with (0,1) uniform x).
    (For two levels of B the values had been mirrored at the mean.)
11. Left skewed (transformation $\log_2(1+x)$ with (0,1) uniform x) with unequal variances on B with a ratio $\max(s_j^2)/\min(s_j^2)$ of 4 (correlation n_j with s_j^2 r=0.12).
12. Same as 11, but with unequal variances on both factors (correlation n_{ij} with s_{ij}^2 r=0.04).
13. N(0,1) with unequal variances on both factors with a ratio $\max(s_{ij}^2)/\min(s_{ij}^2)$ of 3 for unequal cell counts where small n_{ij} correspond to small variances (*positive pairing,* r=0.98).
14. Same as 13, but with negative pairing (r=0.97).
15. Left skewed (transformation $\log_2(1+x)$ with (0,1) uniform x) with unequal variances on both factors with a ratio $\max(s_{ij}^2)/\min(s_{ij}^2)$ of 3 for unequal cell counts where small n_{ij} correspond to small variances (*positive pairing,* r=0.98).
16. Same as 15, but with negative pairing (r=0.97).

In the cases of heteroscedasticity the cells with larger variances do not depend on the design. Subsequently i,j refer to the levels of factors A respectively B.

- For both designs and unequal variances on B the cells with j=1 have a variance ratio of 4 and those with j=2 a ratio of 2.25.
- For both designs and unequal variances on A and B the cells with i=1 and $j \leq 2$ have a variance ratio of 4 and those with i=2 and $j \leq 2$ a ratio of 2.25.

The main simulation study consists of three parts:

- The type I error rates are studied for a fixed n_{ij} (depending on the design) and fixed effect sizes. For this purpose every situation had been repeated 5000 times. This seems to be the current standard.
- Further the error rates are computed also for n_{ij} varying from 5 to 50 in steps of 5 and for fixed effect sizes (see below), in order to see on one side, if acceptable rates stay acceptable, and on the other side, if too large rates get smaller with larger samples. For the same situations the power rates are computed.
- Additionally the error rates are computed for increasing effect sizes, but fixed n_{ij} (depending on the design), to see the impact of other nonnull effects within a model. The effect sizes are varying from 0.1*s to 1.5*s in steps of 0.2*s (s being the standard deviation of the dv). For the same situations the power rates are computed, but with effect sizes varying from 0.2*s to 0.9*s in steps of 0.1*s .

In contrast to the first part a repetition of 2000 times had been chosen for the computation of the error rates and power for large n_{ij} as well as increasing effect sizes, not only because of the larger amount of required computing time, but also because the main focus had been laid more in the relation between the methods than in exact values. A preliminary comparison of the results for the computation of the power showed that the differences between 2000 and 5000 repetitions are negligible. By means of a unique starting value for the random number generator the results for all situations rely on the same sequence of random numbers and therefore on identical sam-

ples. This should make the results better comparable.

Concerning the graphical representation of the power two graphs have been chosen:

- the absolute power as the proportion of rejections in percent and
- the relative power, which is computed as the absolute power divided by the 25% trimmed mean of the power of the 8 methods for each n_{ij} =5,...,50 or d=0.2*s,...,0.9*s and should make differences visible in the area of small n_{ij} or d where the graphs of the absolute power of the 8 methods lie very close together.

Effect sizes

The main focus had been laid upon the control of the type I error rates for α=0.05 and α=0.01 for the various methods and situations as well as on a comparison of the power for the methods. For the computation of the random variates level/cell means had to be added corresponding to the desired effect sizes. These are denoted by a_i and b_j for the level means of A and B corresponding to effects α_i and β_j , and ab_{ij} for the cell means concerning the interaction corresponding to effects $\alpha_i + \beta_j + \alpha\beta_{ij}$.

For the subsequent specification of the effect sizes the following abbreviations are used (*s* being the standard deviation):

- A(*d*): a_1=*d**s*, a_2=0 for a 2*4 plan,
  respectively a_1= a_2= *d**s*, a_3= a_4= 0 for a 4*5 plan
- B(*d*): b_1= b_2= *d**s*, b_3= b_4= 0 for a 2*4 plan,
  respectively b_1= b_2= *d**s*, b_3= b_4= b_5= 0 for a 4*5 plan
- AB(*d*): ab_{11}= ab_{12}= ab_{23}= ab_{24}= *d*s/2* and ab_{21}= ab_{22}= ab_{13}= ab_{14}= *-d*s/2* for a 2*4 plan,
  respectively ab_{ij}=0 except ab_{11}= ab_{12}= ab_{21}= ab_{22}= ab_{34}= ab_{35}= ab_{44}= ab_{45}= *d*s/2* ,
  ab_{14}= ab_{15}= ab_{24}= ab_{25}= ab_{31}= ab_{32}= ab_{41}= ab_{42}= *-d*s/2* for a 4*5 plan

The error rates had been checked for the following effect models:

- main effects and interaction effect for the case of no effects (null model, equal means) and for the case of one significant main effect A(0.6),
- main effect for the case of a significant interaction AB(0.6) as well as for the case of a significant main effect A(0.6) and interaction AB(0.6),
- interaction effect for the case of both highly significant main effects A(0.8) and B(0.8).

These are 7 models which are analyzed for both a balanced and an unbalanced design. So there are all in all 14 models.

For the power analysis the effect sizes had to be reduced in order to distinguish better the power for cell counts between 20 and 50. The following situations and effect sizes had been chosen:

- power of main effect A(0.3) in case of no other effects, in case of a significant effect B(0.3), in case of a significant interaction AB(0.4) and in case of a full model (B(0.3) and AB(0.4)),
- power of interaction effect AB(0.4) for the case no main effects, for the case of a significant main effect A(0.3) and in case of a full model (A(0.3) and B(0.3)).

Handling right skewed distributions

Rather unproblematic behaves the exponential distribution because it has only one parameter for both mean and variance. So there is no differentiating between the cases of equal and

unequal variances. To analyze the influence of effects *d* it is not reasonable to add a constant *d*s* to the values *x* of one group. In order to keep the exponential distribution type for the alternative hypothesis (H_1) a parameter λ‘ had to be chosen so that the desired mean difference $1/\lambda - 1/\lambda'$ is *d*s* where in this case $s=(1/\lambda + 1/\lambda')$. As a consequence the H_1-distribution has not only a larger mean but also a larger variance.

In contrast the lognormal distribution reveals a more unfriendly behavior: all nonparametric methods under consideration here show increasing type I error rates for increasing sample sizes in the case of heterogeneous variances. The extent differs from the skewness and from the degree of variance heterogeneity. For larger skewed lognormal distributions, e.g. with parameters $\mu=0$ and $\sigma^2=1$, things look a bit different: the ART- and to a less degree also the ART+INT-technique cannot keep the type I error under control even for homogeneous variances and equal cell counts, with rates usually between 8 and 11 percent. A more precise investigation of the error rates of the lognormal distribution has been done recently by Luepsen (2016), who confirmed earlier results by Carletti & Claustriaux (2005) and Zimmerman (1998) as well as studies by G. Vallejo et al. (2010), Keselman et al. (1995) and Tomarken & Serlin (1986). Tables of the type I error rates for the tests of the null model for all methods and various situations are to be found in appendix A 6. As the behavior does not differ essentially for different parameters, a lognormal distribution with parameters $\mu=0$ and $\sigma^2=0.25$ has been chosen for the comparisons here. Its shape resembles slightly the normal distribution with a long tail on the right. Only equal variances will be considered here. As distribution for the alternative hypothesis (H_1) a shift of the distribution of the null hypothesis (as described in the previous section) is one choice, thus keeping equal variances. But with real life right skewed data the distribution of the alternative hypothesis often includes a change both of means and variances. In this case a different lognormal distribution had to be selected for H_1 so that the means have the desired difference, e.g. $\bar{x}$ and $\bar{x}+d^*s$, but slightly different variances. Preliminary tests for the calculation of the power showed that both models produce nearly the same results. Therefore the first method has been preferred because of the easier computational handling.

Additionally another right skewed distribution (above number 4) is included that has a form comparable to the lognormal distribution with parameters $\mu=0$ and $\sigma=0.8$, but restricted to the interval [0.67 , 2], or also comparable to a right shifted exponential distribution. The results for this one comprise no peculiarities and shall not be discussed here.

5. Results I: Type I error rates

Tables and Graphical Illustrations

It is evident that a study considering so many different situations (8 methods, 16 distributions, 2 layouts, and 7 models) produces a large amount of information. Therefore the following remarks represent only a small extract and will concentrate on essential and surprising results. All tables and corresponding graphical illustrations are available online (see address below). These are structured as follows, where each table and graphic includes the results for all 8 methods and report the proportions of rejections of the corresponding null hypothesis:

- appendix A 1: type I error rates for fixed n_{ij} = 5 and n_{ij} =10 and for $\alpha=0.05$, $\alpha=0.01$, equal and unequal cell counts and for different models,
- appendix A 2: type I error rates for large n_{ij} (5 to 50 in steps of 5) for $\alpha=0.05$ and fixed effect sizes, for equal and unequal cell counts and for different models,
- appendix A 3: power in relation to n_{ij} (5 to 50 in steps of 5) referring to $\alpha=0.05$ and fixed effect sizes, for equal and unequal cell counts and for different models,

- appendix A 4: type I error rates for large effect sizes (0.1**s* to 1.5**s* in steps of 0.2**s*) for α=0.05 and fixed n_{ij}, for equal and unequal cell counts and for different models,
- appendix A 5: power in relation to increasing effect sizes from 0.2**s* to 0.9**s* in steps of 0.1**s* for α=0.05 and fixed n_{ij}, for equal and unequal cell counts and for different models,
- appendix A 6: type I error rates for large n_{ij} (5 to 50 in steps of 5) for α=0.05 and fixed effect sizes for various lognormal distributions,
- appendix A 7: type I error rates for small and large n_{ij} (5, 10 and 50) for α=0.05 and fixed effect sizes of the exponential and the uniform distributions, each for the version of a continuous and three versions of a discrete distribution.

All references to these tables and graphics will be referred as A *n.n.n*. The most important tables of A 1 and some graphics of A 2 to A 5 are included in this text. All tables and graphics can be viewed online:
http://www.uni-koeln.de/~luepsen/statistik/texte/comparison-tables/

Criteria

A deviation of 10 percent ($\alpha + 0.1\alpha$) - that is 5.50 percent for α=0.05 - can be regarded as a stringent definition of robustness whereas 25 percent ($\alpha + 0.25\alpha$) - that is 6.25 percent for α=0.05 - to be treated as a moderate robustness (see Peterson, 2002). It should be mentioned that there are other studies in which a deviation of 50 percent, i.e. ($\alpha \mp 0.5\alpha$), Bradleys liberal criterion (see Bradley, 1978), is regarded as robustness. As a large amount of the results concerns the error rates for 10 sample sizes n_{ij} = 5,...,50 it seems reasonable to allow a couple of exceedances within this range.

Performance for large *n:* general behavior of the different methods

Whereas there are no spectacular results for small sample sizes n_{ij} (see chapter 8 for details) things look quite different for increasing cell counts. Only the parametric F-test shows no differences in the behavior between small and large sample sizes n_{ij} . Perhaps to mention: Exceeding error rates decrease often with increasing n_i (see e.g. A 2.2.12, A 2.4.3 and A 2.6.12) which had to be expected from the central limit theorem. Insofar the results confirm the attributes of the F-test mentioned in chapter 1. Elsewise the nonparametric procedures. For larger n_{ij} some show rising error rates, especially the ART, ART+INT, RT, ATS and sometimes the INT and the Puri & Sen procedures. The following peculiarities do neither concern those unbalanced designs where n_{ij} are correlated with s_{ij} nor discrete distributions that will be looked at later.

Generally the ART method tends to be liberal with rates above the acceptable limit of moderate robustness (beyond 6) in the cases of heterogeneous variances (see e.g. figure 3). Additionally for the test of a main effect in an unbalanced design, if there are other nonnull effects, the error rates rise to 10 and above when n_{ij} ($n_{ij} > 15$) increases up to 50 for all distributions (see figure 1 as well as A 2.4, 2.6 and 2.8). The ART+INT shows a similar performance as the ART which is plausible from the procedure, especially with the same deficiencies for the tests of main effects in unbalanced designs. But mostly its rates lie below those of the ART as remarked by Carletti & Claustriaux (2005). Additionally there are several settings of heterogeneous variances where the ART+INT, on the contrary to the ART, keeps the error rate completely under control: e.g. all tests of main effects (see A 2.1 and A 2.2), except the case noted above. And if there are deficiencies with larger n_{ij} they start normally with $n_{ij} > 15$.

A striking phenomenon is the behavior of the ATS and the RT in most situations of unequal variances for the tests for main and interaction effects when there is another nonnull effect: the

error rates increase up to 10 and sometimes above when n_{ij} increases up to 50 (see figure 2 and see sections 2 and 3 as well as 11 an 12 in A 2.3 to A 2.8 and A 2.11 to A 2.14). But it has to be remarked that they stay in the acceptable region for $n_{ij} < 15$. This is the phenomenon described in chapter 2 for RT, but happening here only in the case of unequal variances. The same applies to the Puri & Sen-method, but its rates lie clearly below those of the RT and ATS, but with values between 6 and 10 still beyond the acceptable limit. The conservative behavior was explained in chapter 2. So the Puri & Sen-method keeps the type I error rate often in the moderate robustness interval, frequently even the stringent robustness interval at least for small and moderate $n_{ij} < 25$, in situations where the RT exceeds the limits (see e.g. A 2.6.3, 2.7.3, 2.7.11, 2.11.12 or 2.13.2). If the Puri & Sen-method offends the criterion then only for larger $n_{ij} \geq 30$.

The INT-procedure has also some problems with unequal variances but predominantly in unbalanced designs showing slightly increased error rates between 7 and 10 (see e.g. A 2.4.11, 2.10.12 and 2.13.12). Additionally the rates rise above the limit in a couple of cases with equal variances but underlying skewed distributions (see A 2.4.10, 2.7.4, 2.8.4 and A 2.14.4). And finally the behavior seems to be generally slightly liberal for the test of the interaction if both other effects are nonnull (see A 2.13 and A 2.14).

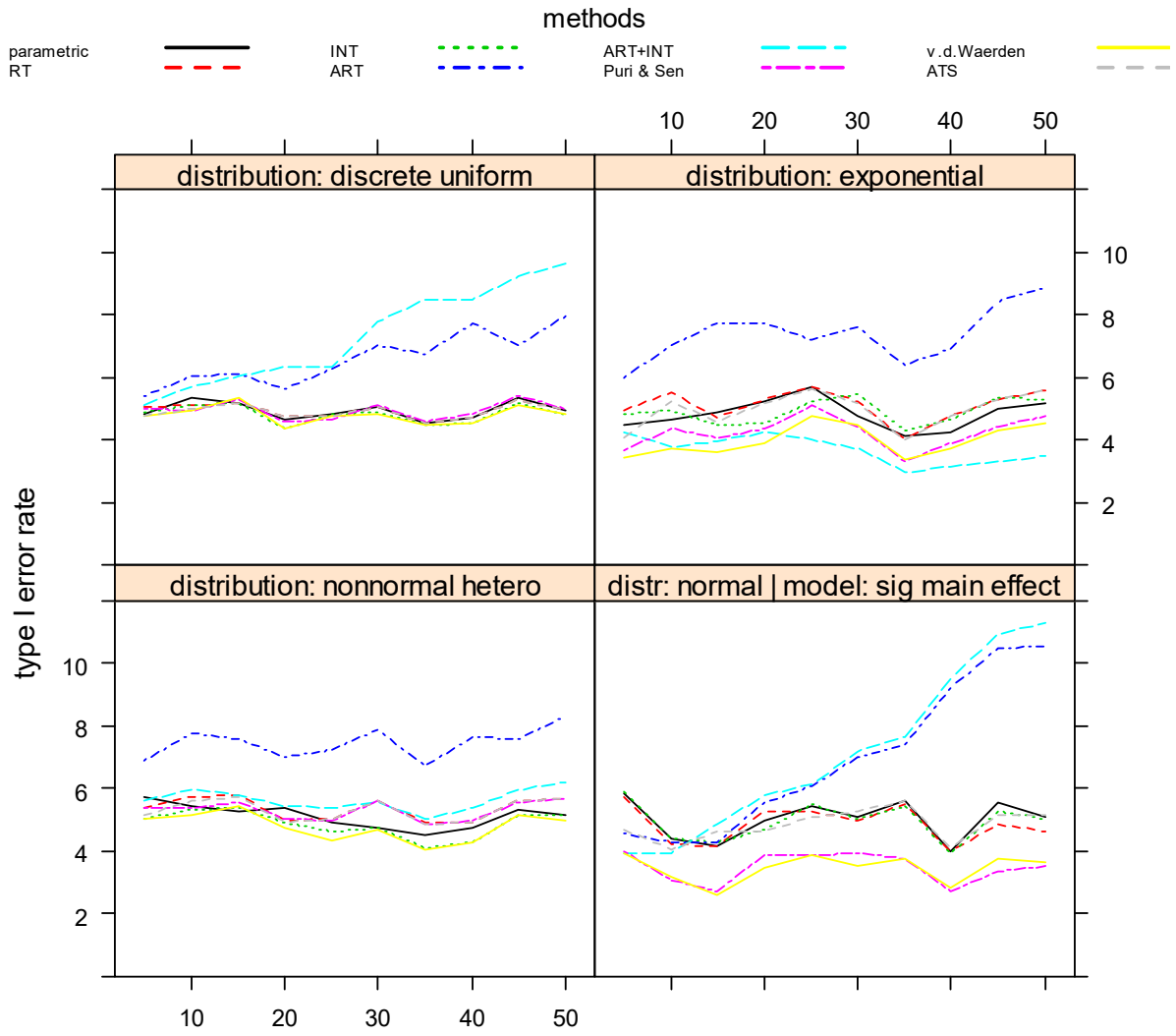

Figure 1: type I error rate (main effect) in four situations where the ART and ART+INT fail: underlying discrete uniform distribution, continuous exponential distribution, nonnormal distribution with heterogeneous variances (all in balanced designs) and when there is another significant main effect (in a unbalanced design). (The cell counts are on the base line.)

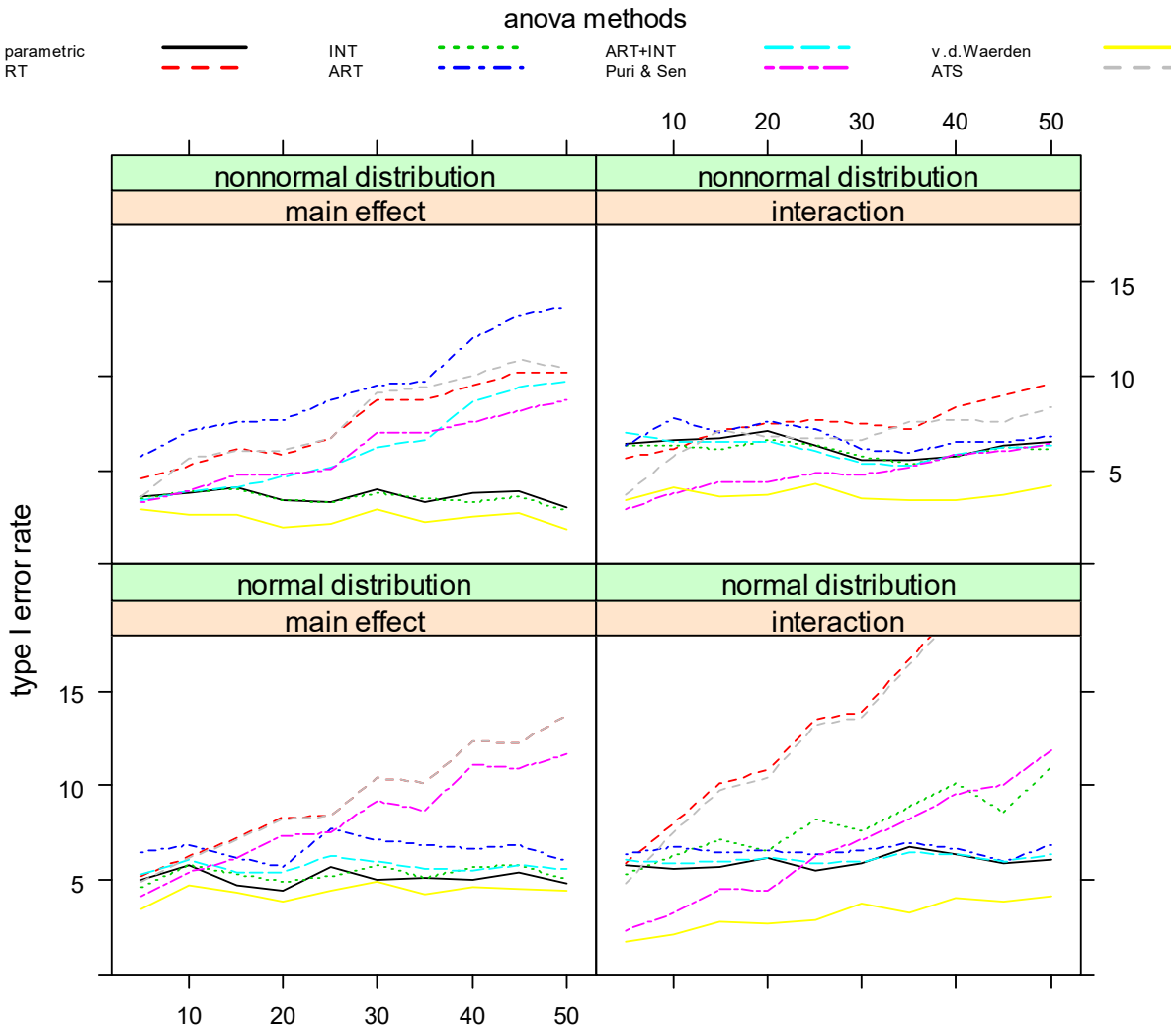

Figure 2: Four situations with unequal variances where the type I error rates for the RT, Puri & Sen and the ATS exceed the acceptable range with increasing cell counts: test of main and interaction effect in the case of a nonnormal distribution and unequal n_{ij} as well as in the case of a normal distribution and equal n_{ij} . (The cell counts are on the base line.) (Note that here the van der Waerden as well as the parametric F test behave benevolently.)

The van der Waerden-test is the less conspicuous from all methods. The shape of the graph of its rates looks much alike them of the INT-method, which is not surprising considering the computation, but the values lie clearly lower. So only three singular instances exist where the test reacts slightly liberal with error rates between 6 and 7 (see A 2.2.12, 2.8.5 and 2.10.12).

Unequal variances s_{ij}^2 together with unequal sample sizes n_{ij}

When s_{ij}^2 and n_{ij} are uncorrelated the van der Waerden test is the only one that has the type I error under complete control, e.g. for the interaction where the rates of the parametric F-test lie between 7 and 12 percent (see figure 3 and sections 3 and 12 in A 2.10, 2.12 and 2.14). Perhaps to mention that in case of the normal distribution for n_{ij}~10 90 percent of the sample correlations were located in the interval [-0.29 , 0.28] and for n_{ij}~50 in [-0.19 , 0.08]. In the case of positive pairing nearly all methods show acceptable type I error rates, except the ART and ART+INT in the cases of other nonnull effects as noted before. In the challenging case of negative pairing the ATS-method is the only one that keeps the error level under control for all models.

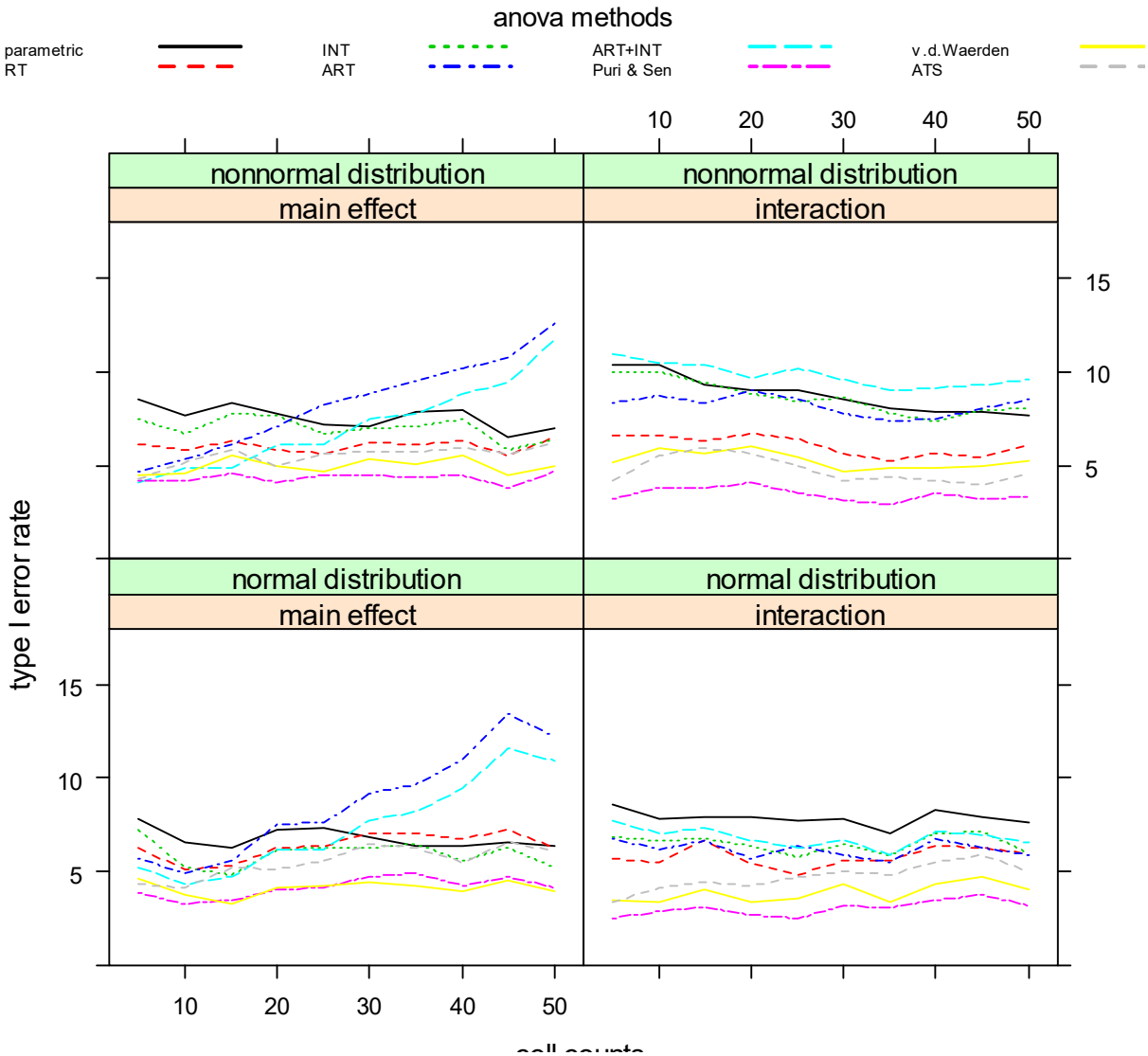

Figure 3: Two situations with unequal variances in unbalanced designs where the type I error rates for many methods including the parametric F-test exceed the acceptable limit: test of main and interaction effect in a model with another significant main effect in the case of a nonnormal (left skewed) distribution as well as in the case of a normal distribution. (Note that here the van der Waerden test behaves benevolently.)

Performance for large *n*: right skewed distributions

For the exponential distribution it has to be remarked that in most situations the type I error rates of the ART-procedure rise beyond the acceptable limit for n_i larger than 10 or 20 (see figure 4 and e.g. A 2.3.5, 2.4.5, 2.6.5 and 2.8.5 with values between 9 and 20), except for the tests of the null model. And the ART performs even worse in the version with integer values. This phenomenon had been analyzed in detail and explained by Luepsen (2017). As a consequence the same applies also to the ART+INT-procedure, but to a less degree. Additionally there are a couple of situations where the RT, Puri & Sen and ATS react liberal: for the test of a main or interaction effect if both other effects are nonnull (see figure 4 and A 2.7.5, 2.8.5, 2.13.5 and 2.14.5).

For the lognormal distribution equal variances have been assumed in this study, as mentioned in chapter 4. Therefore it is not surprising that the error rates do not show any peculiarities.

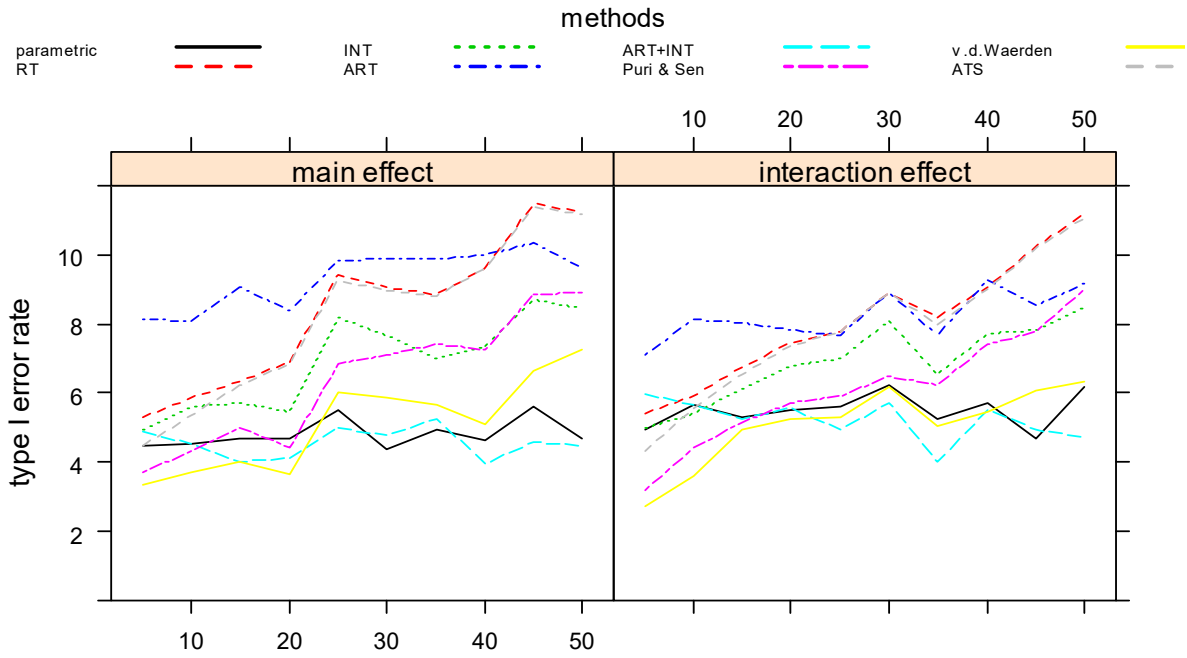

Figure 4: Two situations with an underlying exponential distribution where the type I error rates for the RT, Puri & Sen, ATS and ART exceed the acceptable range with increasing cell counts: test of a main effect when the other main effect is significant and test of interaction effect when both main effects are significant, both for the case of equal n_{ij}. Note that the ART+INT keeps the error under control in these situations in contrary to the ART.

Discrete Variables

Though all the nonparametric procedures under consideration here, except the ATS, require a continuous dependent variable, in practice they are applied to discrete variables as well and often even to ordinal variables with only a few distinct values. A comparison of all 8 methods with regard to the behavior in the case of underlying discrete distributions, exponential and uniform, shows that the type I error rates rise mainly for the ART- and the ART+INT-procedures for increasing cell counts n_{ij} up to 20 percent (more details in chapter 8). An extensive study about the impact of discrete dependent variables comes from Luepsen (2017) in which also an explanation of this phenomenon is given. Additionally it is shown that the error rates rise beyond the interval of moderate robustness if the number of distinct values decreases, and this more severe for the exponential than for the uniform distribution.

Impact of effect size on the Type I error rates

At first sight the overall decreasing error rates of the Puri & Sen and the van der Waerden methods for rising effect sizes are remarkable (see e.g. A 4.1.1 and 4.6.1). But this was explained in chapter 2. The ART and the ART+INT show unacceptable error rates for growing effect sizes similar to those described for large *n* in the previous section: underlying exponential distribution (see e.g. A 4.1.5, 4.2.5, 4.3.5, 4.4.6 etc.), test of main effects in unbalanced designs if there are also other nonnull effects in the model (see A 4.2 and A 4.4), and additionally the ART for underlying discrete distributions. Finally a look on the performance of the RT since it is said to show increasing error rates for the interaction if there are also significant main effects (see section 2.1). There is only a slight increase to observe with maximum rates between 7 and 10 percent for effect sizes of 1.5*s* in the cases of heterogeneous variances (see sections 2 and 3 as well as 11 and 12 in A 4.5 and A 4.7) and in the case of the exponential distribution (see A

4.7.5), both in balanced designs. And as stated by Huang (2007), the INT keeps the error rate under control in most of these situations. It remains to remark that the ATS-method performs rather similar to the RT-method. All other methods behave inconspicuously.

Summary

effect (model)	des	param	RT	INT	ART	ART+ INT	Puri & Sen	van der Waerden	ATS
A	eq				4 **6** 8 **9** **B** **C**	**6** **9**			
(null model)	ne	3 C		C	**6**	**6** **9**		C	
B	eq				**5** **6** **B** **C**	**A**			
(A sig)	ne	**3** **C**	**B**	**A** **C**	**1...6** **8...C**	**1...6** **8...C**	B		**B**
A	eq		**2** **3** **B** **C**		2 3 **5** **6** 9 B C	**6** **9**	**2** **3** **B** **C**		**2** **3** **B** **C**
(AB sig)	ne	3 5 6 C	**2** **3** **B**		**1....C**	**1....C**	**2** **B**		**2** **B**
B	eq		**3** **5** **6**	4 5	**5** **6** **B** **C**	**A**	**5** **6**		**5** **6**
(A+AB sig)	ne	3 **C**	**3** **5** **6** **C**	**2** **4** **5** **6** C	**1....C**	**1....C**	**3** **5** **6** **C**	**5** 6	**3** **5** **6** **C**
AB	eq	2 B		B C	2 3 6 B C	2 C			
(null model)	ne	**3** B **C**	C	3 B **C**	2 3 **6** **9** **B** **C**	**3** **6** **9** B **C**	C	**C**	
AB	eq	B	**2** **3** **B** **C**	2	2 3 **5** **6** B C	2 C	**2** **3** **B**		**2** **3** **B** **C**
(A sig)	ne	**3** **C**	**2** 4 **B**	3 4 **C**	**3** 6 **9** B **C**	**3** **9** **C**			**2** **B**
AB	eq	B	**2** **3** **5** **6** **B** **C**	2 **3** 4 5 6 C	**5** **6** B C		**3** **5** **6** **C**		**2** **3** **5** **6** B C
(A+B sig)	ne	**3** B **C**	**2** **3** 4 **5** **6** **B** **C**	**3** **4** 5 6 B **C**	2 3 6 **9** **B** **C**	**3** **9** **C**	**C**		**2** **3** **5** **6** **B** **C**

Table 1: Violations of the type I error rates in the range of n_{ij} = 5,...,50
The numbers refer to the distributions (see chapter 4.), A to right/left-skewed distributions, B and C to left skewed distributions with unequal variances. The layout has the following meaning:
n: moderate - values outside the interval of moderate robustness, but mostly below 7
n: strong - nearly all values above 7
n: rising - values inside the interval of moderate robustness for $n_{ij} < 15$, but rising for larger n_{ij}
„eq“ and „ne“ in the column „des“ refer to equal and unequal cell counts.

The ART- and the ART+INT-procedures have deficiencies with heterogeneous variances, with discrete variables, with (even slightly) right skewed distributions and with the test of main effects in unbalanced designs. This makes these methods not recommendable though the ART+INT shows acceptable rates at least for $n_{ij} < 15$. And the positive results mentioned in chapter 3 are not valid in general. The RT-, ATS- and Puri & Sen-method have generally problems with unequal variances, even for balanced designs. And these problems enlarge for tests in those cases when there are other nonnull effects. On the other side the ATS is the only method that can handle in all situations the challenging case of unbalanced designs with unequal variances where small n_i correspond to large s_i. But also for the ATS it must be admitted that the control of the type I error rate cited in chapter 3 is no more valid for larger samples. The INT-method is in many cases acceptable though there are a number of unsatisfying situations for which there is no guideline visible. From this it is obvious that the van der Waerden-test has the

fewest violations, especially for heterogeneous variances. Table 1 gives an impression of the distribution of error rates offending the limits for the different situations.

6. Results II: Power

In this study only the relation between the powers of the different nonparametric anova methods is examined whereas the absolute power values achieved are of minor interest. The results for equal and unequal cell counts are only conditionally comparable because of the different number of cells as well as the different cell counts. From the previous section it is obvious that, besides the van der Waerden-test, the nonparametric methods are scarcely able to achieve amelioration for the cases of unequal cell frequencies paired with unequal variances, compared with the parametric F-test. Therefore the focus is laid here on those settings with non-normal distributions where nonparametric methods are supposed to reach a higher power than the parametric F-test. Of course there are situations in which tests react liberal, leading on one side to high power rates, but also on the other side to offending the type I error rate. Such situations will be neglected here.

Performance of the different methods

Table 2 gives an impression of the distribution of above-average power performances. For every sample size a performance value - in the graphics of appendix A 3 denoted by relative power - is computed as the percentage of power above the mean over the 8 methods, which is computed as a 25% trimmed mean. These values are averaged over all sample sizes n_{ij} = 5,...,50 as well as for small sizes n_{ij} = 5,...,20 and for large sizes n_{ij} = 25,...,50. This table demonstrates among others the poor performance of the Puri & Sen- and the ATS-methods which never show values that lie 5 percent above the average. Further it shows that the power of the v.d.Waerden-test shrinks when there are side effects, and of course the good performance of the INT- and especially of the ART+INT-procedure.

As a general result, considering all forms of distribution and effect situations, it can be concluded that the methods based on the inverse normal transformation (INT, ART+INT and v.d. Waerden) show constantly a high power, and in many cases even the best power (see figures 5a and 5b), especially for non-normal distributions, with one exception: the case of exponential distributions. The ART+INT performs best when there are also other effects present, in contrary to the v.d.Waerden-method, which complies better if there are no other significant effects. And this applies in a boosted degree to the interaction effect. The superiority of both methods starts in most cases only with $n_{ij} > 10$. But unfortunately the ART+INT shows unacceptable type I error rates in many of these situations. More details in chapter 8. The INT and v.d. Waerden methods are the best for the power of main effects as well as for underlying uniform or right skewed distributions (see figure 5a and table 2 as well as A 3.7, A 3.8, A 3.13 and A 3.14). There are no essential differences, neither between the balanced and the unbalanced design.

The ART-procedure shows high power rates only in situations where it shows a liberal reaction for the type I error. In all other cases, e.g. for underlying uniform distributions, its power is rather poor. Also for the RT holds: the good performance occurs in those situations where the error rates exceed the limits, worsens when there are also other effects and is rather poor for the full model. And the ATS and Puri & Sen methods which keep the type I error rate the best in many cases? In general these are among those with the lowest power. Table 2 demonstrates that these have never an above-average power, frequently the lowest power, e.g. for the main and interaction effect in the full model (see e.g. A 3.7 and 3.8, 3.11, 3.13). For the Puri & Sen method this effect is plausible because the reduction of error sum of squares induced by significant main

effects cannot be exploited by the Puri & Sen method. This applies also to the van der Waerden method but in that case this negative effect is compensated by the normal transformation. The ATS is often the worst performer with power rates about 40 percent below average (see e.g. A 3.10 and 3.12). Nevertheless there are a few situations in which the ATS excels positively: in unbalanced designs with heterogeneous variances if large n_{ij} correspond to large s_{ij} (see A 3.2.13, 3.2.15, 3.10.13, 3.10.15, 3.14.13 and 3.14.15). And what about the power of the parametric F-test? In general its power lies in the middle of the results. However the excellent performance for underlying exponential distributions should be mentioned here. Details in chapter 8.

effect (side effects)	des	param	RT	INT	ART	ART+ INT	Puri Sen	van der Waerden	ATS
A	eq	**5 6**		**4** 7 **8 9 A** B	2 3 C	4 8 9 A B		**4 8 9 A** B	
(none)	ne	**5 6**	2 4 9	1 **4** 7 8 **9 A** B C	2	8 A		**4** 8 **9 A** C	
A	eq	**5 6**		4 7 **8 9 A** B	2 3 C 5 6	3 4 **8 9 A** B C		**4 8 9** A	
(B sig)	ne	**5 6**	2 4	1 3 **4** 7 **8 9 A** B C		**9**		**4 8 9 A** C	
A	eq	**2 3 5 6** A **B C**		**4 8 9 A B** C	**2 3** 5 6 **B C**	**2 3** 4 **8...C**		**4 8** 9...C	
(AB sig)	ne	1 2 3 **5 6** A B C		**4** 5 6 7 **8 9 A** B C	**2 3** 5 6 **B C**	1 **2** 3 4 **8...C**		**4 8** 9 A B	
A	eq	2 **3 5 6** B A **C**		4 **8 9** A B	**2 3 5 6** B **C**	**2...4** 5 6 7 **8...C**		8 9 B	
(B+AB sig)	ne	1 **2 3 5 6** A **B C**	4 7	1 4 7 **8..A B** C	1 **2 3** 4 **5 6 B C**	1 **2 3** 4...C		8 9 B	
AB	eq	**5 6**		4 6 **8 9 A** B	2 3 5 **6** C	4 **8 9 A** B		4 **8 9** A	
(none)	ne	3 **5 6**	2	**4** 6 7 **8 9 A** B C	2 3 5 6 B	1 4 **8 9 A** B C		4 **8 9 A**	
AB	eq	1 2 3 **5 6** B C		**4** 5 6 7 **8 9 A** B C	**2 3** 5 6 B C	**2 3** 4 **8 9 A** B C		**4** 6 **8** 9..C	
(A sig)	ne	1 2 **3 5 6** 9...C		1 2 3 **4...A** B C	2 3 5 6 B C	1 **2 3 4** 5 7 **8..C**		4 5.. B	
AB	eq	1 **2 3 5 6** A **B C**		1...5 **6...B** C	**2 3** 4 **5 6** B **C**	1 **2...4** 5 6 **7...C**		6 8.. B	
(A+B sig)	ne	1 **2 3 5 6 A B** C	1 4 7	1 2 3 **4...B** C	**2 3** 4 **5 6** 7 B **C**	1 **2...4** 7 **8...C**		6 8.. C	

Table 2: Above-average power performance in the range of n_i = 5,...,50
The numbers refer to the distributions (see chapter 4), A (10) to right/left-skewed distributions, B (11) and C (12) to left skewed distributions with unequal variances. The layout has the following meaning:
n: moderate - power at least 5 percent above the average
$\overline{n}$: moderate - power at least 5 percent above the average, only for large samples (above n_{ij} >20)
$\underline{n}$: moderate - power at least 5 percent above the average, only for small samples (below n_{ij} <20)
n: strong - power at least 10 percent above the average
„eq“ and „ne“ in the column „des“ refer to equal and unequal cell counts.

Variances s_i^2 correlated with sample sizes n_i

If small n_{ij} correspond to small s_{ij} the ATS is the best performer (see e.g. A 3.2.13 and 3.2.15). If not available the ART as well as the RT method can be applied. If small n_{ij} correspond to large s_{ij} the high values of the parametric F-test and the INT-method are not helpful because of their poor type I error control. Therefore the v.d.Waerden-procedure is the only suitable one showing good results (see e.g. A 3.2.14 and A 3.2.16), especially for skewed distributions.

Impact of effect size on the Power

In general the results for varying effect sizes are similar to those for varying cell counts. First it must be remarked that the power for all tests (main and interaction effects) in unbalanced models is lower than in balanced designs, though the smaller n_{ij} is counterbalanced by the larger number of cells. And this applies particularly to the tests in the full model.

Whereas the van der Waerden-method showed generally a good performance concerning the power in relation to the sample size this applies not in the same extent to the impact of the effect sizes. The power rises more with increasing n_i than with increasing effect sizes. When there are other nonnull effects the power rates of the van der Waerden-method lie often below the mean (see e.g. A 5.5, 5.6, 5.9 and 5.10). This can be explained by the computational process. So it is not surprising that this applies also to the Puri & Sen-method. However thanks to the INT-trans-

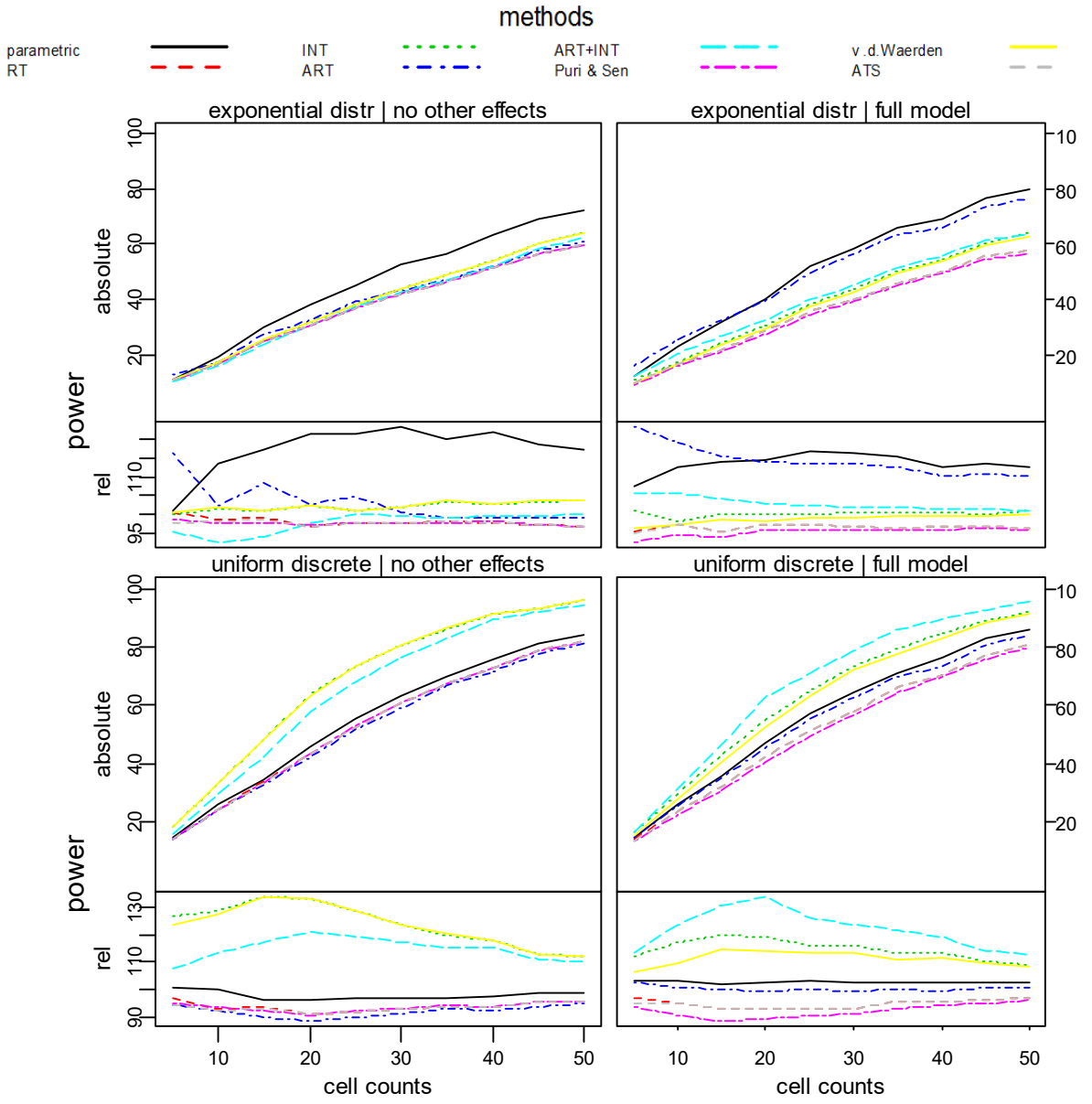

Figure 5a: Power of a main effect in a balanced model with and without other nonnull effects with two different underlying nonnormal distributions: exponential continuous and uniform discrete. The first row shows the excellent performance of the parametric F-test, the second the decreasing power of the v.d.Waerden test if there are other effects present.

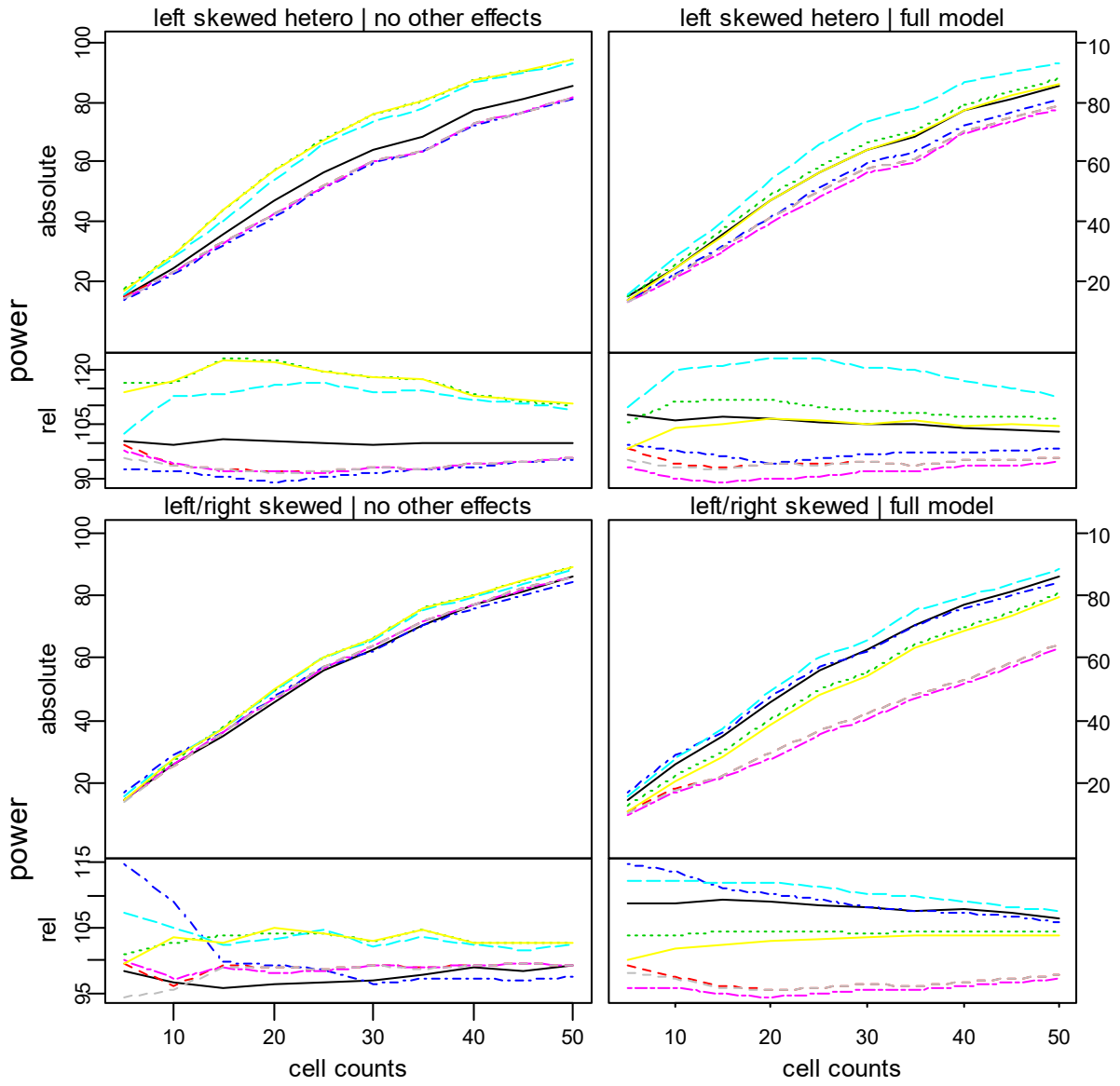

Figure 5b: Power of a main effect in a balanced model with and without other nonnull effects with two different underlying nonnormal continuous distributions: left/right skewed and left skewed with heterogeneous variances.

formation the v.d.Waerden rates are slightly better than those of the Puri & Sen-test. In the worst case (interaction effect of the full model in an unbalanced design, see A 5.10) the rates reach just 50 percent of the best performing method. Nevertheless the van der Waerden-method performs generally well for nonnormal distributions, especially for the uniform, right skewed and the mixture of skewed distributions.

Summary

Summarizing the results so far: The van der Waerden-method is generally among the good performers, though it weakens for small n_{ij} or if there are also other nonnull effects. Just in these situations the ART+INT-procedure is recommendable, especially for heteroscedastic normal distributions and for lognormal distributions. The INT-method is a good choice for all right skewed and uniform distributions as long as there are no heterogeneous variances. The ART reveals a good performance just in those cases where its type I error behavior is unsatisfactory.

And finally the parametric F-test reaches a high power for the exponential distribution, and of course for normal distributions, but in the case of unequal variances only for balanced designs.

7. Conclusion

From the previous sections it is obvious that there is no one best method. While the van der Waerden-test seems to be the best one with regard to the type I error rate there is not a single best method concerning the power. Nevertheless, if not much testing is desired the van der Waerden-method seems to be the first choice because it has been attested a good type I error behavior and shows generally a good power performance. But with the following constraint: the van der Waerden method has a reduced power for small n_{ij} or if there are other nonnull effects, e.g. in the case of a full model. For these situations the ART+INT can be chosen as an alternative, especially for the test of interaction effects: on one side its type I error violations are moderate and most deficiencies occur mainly for larger cell sizes $n_{ij} \geq 20$, and on the other side this method has a large power as stated above.

The pure ART-technique is generally not recommendable. Instead the ART+INT is preferable due to the dampening effect of the INT-transformation on the error rates and due to the better power performance. As the ART-techniques violate the type I error rate for the test of main effects in unbalanced designs the INT-method should be preferred for $n_{ij} \geq 20$ in these cases. The rather simply computable RT-method appeared not as bad as it is often described. Only in the cases of unequal variances it cannot keep the error rates under control. Here also this deficiency occurs mainly for larger cell sizes $n_{ij} > 15$ and for larger effect sizes. But nevertheless, the also easy computable INT-method is preferable since it keeps the error rate better under control than the RT and possesses the far better power than the RT. The ATS- as well as the Puri & Sen-method are not advisable because of their low power. Concerning the type I error control the ATS has the same problems with heterogeneous variances as the RT, whereas the Puri & Sen-test is less problematic.

And the parametric F-test? As long as the variances are equal it is no bad choice. Even in the case of heteroscedasticity the test is still valid when the sample sizes are equal. The error rates are always under control and the power rates lie in the middle. But they are inferior to those of the INT-based procedures in the cases of nonnormal distributions. And finally, the parametric F-test is the first choice in the case of right skewed distributions: it is the only one with more or less acceptable error rates in the case of right skewed distributions with slightly unequal variances (see A 6) and it has the best power for underlying exponential distributions.

Some final remarks on the case when unequal cell counts are paired with unequal variances. The case of positive pairing is rather unproblematic because for all methods except the ART the type I error rate lies even in the interval of stringent robustness. In the case of negative pairing the ATS is the only method that keeps the error level under control in every situation. The corresponding low power, between 1/3 and 1/2 of that of the other methods, has to be accepted. When s_{ij}^2 and n_{ij} are uncorrelated the van der Waerden-test is the only one that has the type I error under complete control and shows satisfying power rates.

8. Additional Results

Here some results are reported which mainly confirm previous findings from other articles.

Type I error rates for small *n*

A look onto the results for fixed n_{ij} = 5 and n_{ij} =10 (appendix A 1). For the parametric F-test all the well-known results denoted in the introduction could be confirmed. Concerning the other methods there are also no spectacular results. In the null model (tables 1-1-1 and 1-2-1 in A 1) the ART and ART+INT show only decent exceedances of the moderate robustness in the case of unequal variances. Here applying the INT to the ART shows a dampening effect as already remarked by Carletti & Claustriaux (2005). And in the challenging case of an unbalanced design where small n_{ij} are paired with large s_{ij} only the ATS keeps the error level under control, whereas in the case where small n_{ij} are paired with small s_{ij} of course all tests show acceptable rates (table 1-2-2 in A 1).

When there is a nonnull main effect (tables 1-3-1 and 1-4-1 in A 1 for balanced designs and table 1-4-3 in A 1 for unbalanced designs) again for the ART and ART+INT the rates of main and interaction effect exceed the interval of moderate robustness in the case of unequal variances and for exponential distributions. Here also the ART+INT has the lower values. The same applies to the case of both nonnull main effects. But in these situations of nonnull effects also the RT and ATS show sometimes too large error rates for unequal variances.

Similar results were obtained at the 1 percent level though results at that level tend to be more liberal in general.

Impact of discrete variables on type I error rates

Comparing all 8 methods with regard to the behavior in the case of underlying discrete distributions, exponential and uniform, the tables and graphics in appendix A 2 show that the type I error rates rise mainly for the ART- and the ART+INT-procedures for increasing cell counts n_{ij} , in most cases beyond 10 percent, but sometimes even up to 20 percent. See e.g. A 2.5.6, 2.5.9, 2.10.6, 2.10.9, situations where the rates remain in the interval of moderate robustness for the corresponding continuous distribution. In the case of the uniform distribution the situation is more transparent because for the continuous distribution the error rates of the ART- and the ART+INT-procedures are almost always under control while for the discrete distribution the rates rise up to values between 6 and 8 if n_{ij} increases up to 50. But it has to be noted that at least for equal cell counts the rates keep acceptable for most models. For details see summary tables A 7.8.3 (exponential distribution) and A 7.8.4 (uniform distribution) which represents a summary of the results for the ART-method tabulated in A 2. On the contrary all other methods behave mostly in the normal range.

Power performance for non-normal distributions

For underlying exponential distributions, both in the continuous and in the discrete version, the parametric F-test is without restrictions the best performer, and especially for unbalanced designs the INT- and the v.d.Waerden-procedures are often also a good choice. For the lognormal distribution the differences between the power rates of the different methods are generally rather small. But in most situations the INT- and the v.d.Waerden-procedures are the leader, followed by the ART+INT-technique. For the uniform distributions, both in the continuous and in the discrete version, the methods based on the inverse normal transformation (INT, ART+INT and v.d. Waerden) show constantly the best power. And the differences between these are minimal. The parametric F-test lies generally below the INT-based methods in the medium range, while all other procedures show comparatively low power rates and reach often only 60 to 70 percent of the top values (see e.g. A 3.10.8 and 3.14.8). Also for the case of mixed left/right-skewed distributions the INT-based methods have the highest power rates, followed by the parametric

F-test (see figure 5b). Again for left skewed distributions with heterogeneous variances the INT-based methods are among the best performers. Unfortunately the INT- as well as the ART+INT-method show also increased error rates for this kind of distributions, at least for unbalanced designs. So the only recommendable procedure left is the van der Waerden-test. It remains to remark that the differences between the power rates are generally small.

Power performance of the parametric F test

In general the power of the parametric F test lies in the middle of the results, except for a few situations: In the ideal case of an underlying normal distribution with homogeneous variances the F-test is of course the best performer though the lead to the nonparametric methods is negligible. In models with more than one significant effect, e.g. the full model, the F-test is able to score (see table 2). In the case of the right skewed distributions the parametric F-test is the absolute winner for the much skewed exponential distributions. On the other side the power of the F-test is among the lowest for the lognormal distribution, especially for larger n_{ij} , though the differences between the rates are fairly small. Table 2 also demonstrates that for this type and for uniform distributions the F-test is always inferior to the INT-based methods.

Power in special situations

Now a glance shall be put on some of the situations concerning the distributions and effect combinations. In the various cases of underlying normal distributions the differences between the methods are rather small as long as there are no influences by other effects. In the case of a full model, i.e. all effects assumed significant, the differences rise up to about 30 percent (see e.g. A 3.8.2, 3.8.3 as well as 3.13.3 and 3.14.3). Generally the ART+INT-method yields high power rates in the case of a full model, both for the main and the interaction effects. But this is only helpful for small sample sizes because for the main effect in unbalanced designs its error rates are not under control if $n_{ij} \geq 20$.

Of special interest is the case of unequal variances where, as stated in the chapter above, nearly all methods suffer from unacceptable type I error rates, but might differ with regard to the power. First the case of an underlying normal distribution. The only methods not strongly affected by the heteroscedasticity are the Puri & Sen- and the v.d.Waerden-tests generally, the ART+INT-technique except for the test of interaction effects in unbalanced designs, and the parametric F-test for balanced designs. From these the v.d.Waerden-method has a slightly better overall power performance than the Puri & Sen-method (see. e.g. A 3.5.3 and 3.8.2). However the ART+INT in general as well as the parametric F-test for the cases of equal sample sizes reach often higher power rates than the other two methods, especially in models with nonnull side effects, e.g. the full model (see table 2 as well as e.g. A 3.7.2 and A 3.7.3). Second the case of non-normal distributions. Here was indicated above that the v.d.Waerden-test is to prefer.

Impact of effect size on the Power

Concerning the impact of the effect size the ART-technique shows exactly the same behavior as previously described for increasing n_{ij}. In contrast the ART+INT performs considerably better. A look at the results for main and interaction effects in full models shows that the ART+INT is a good choice for balanced designs whereas the INT is preferable for unbalanced designs. Disappointing is also the general bad output of the ATS. The ATS shows strength only in the cases of unequal cell sizes where small n_{ij} correspond to large s_{ij} (see sections 10 and 12 in A 5.2, A 5.4, A 5.6, A 5.8 and A 5.10).

9. Software

This study has been programmed in R (version 3.0.2 and later version 3.2.2), using mainly the standard anova function `aov` in combination with `drop1` to receive type III sum of squares estimates in the case of unequal cell counts. For the ART, ATS, factorial Puri & Sen and van der Waerden methods own functions had been written (see Luepsen, 2014). It should be noted that meanwhile there exists the R package `ARTool` (Kay & Wobbrock, 2015) for the ART method and the package `rankFD` for the ATS statistic used here. All the computations had been performed on a Windows notebook.

10. Literature

Akritas, M.G., Arnold, S.F., Brunner, E. (1997). Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs, *Journal of the American Statistical Association,* Volume 92, Issue 437, pp 258-265.

Beasley, T.M., Zumbo, B.D. (2009). Aligned Rank Tests for Interactions in Split- Plot Designs: Distributional Assumptions and Stochastic Heterogeneity, *Journal of Modern Applied Statistical Methods*, Vol 8, No 1, pp 16-50.

Beasley, T.M., Erickson, S., Allison, D.B. (2009). Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behavioral Genetics,* 39 (5), pp 380-395.

Bennett, B.M. (1968). Rank-order tests of linear hypotheses*, J. of Stat . Society* B 30, pp 483- 489.

Blair, R.C., Sawilowsky, S.S., Higgins, J.J. (1987). Limitations of the rank transform statistic, *Communications in Statistics*, B 16, pp 1133-45.

Box, G. E. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics,* 35(2), 290-302.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, pp 144-152.

Brunner, E., Dette, H. and Munk, A. (1997). Box-type approximations in nonparametric factorial designs, *Journal of the American Statistical Association*, 92, pp 1494-1502.

Brunner, E., Puri, M.L. (2002). A class of rank-score tests in factorial designs. *Journal of Statistical Planning and Inference,* 103, pp 331–360.

Brunner, E., Munzel, U. (2002). *Nichtparametrische Datenanalyse - unverbundene Stichproben*, Springer, Berlin.

Carletti, I. , Claustriaux, J.J. (2005). Anova or Aligned Rank Transform Methods: Which one use when Assumptions are not fulfilled ? *Buletinul USAMV-CN*, nr. 62/2005 and below, ISSN, pp 1454-2382.

Conover, W. J. & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician* 35 (3): pp 124–129.

Danbaba, A. (2009). *A Study of Robustness of Validity and Efficiency of Rank Tests in AMMI and Two-Way ANOVA Tests*. Thesis, University of Ilorin, Nigeria.

Danbaba, A. (2012). Comparison of a Class of Rank-Score Tests in Two-Factors Designs. *Nigerian Journal of Basic and Applied Science*, 20(4), pp 305-314.

Dijkstra, J. B. (1987). Analysis of means in some non-standard situations. Technische Universiteit, Eindhoven DOI: 10.6100/IR272914.

Erceg-Hurn, D.M. and Mirosevich, V.M. (2008). Modern robust statistical methods, *American Psychologist,* Vol. 63, No. 7, pp 591–601.

Fan, W. (2006). *Robust means modelling: An Alternative to Hypothesis Testing of Mean Equality in Between-subject Designs under Variance Heterogeneity and Nonnormality*, Dissertation, University of Maryland.

Feir, B.J., Toothaker, L.E. (1974). *The ANOVA F-Test Versus the Kruskal-WallisTest: A Robustness Study*. Paper presented at the 59th Annual Meeting of the American Educational Research Association in Chicago, IL.

Field, A. (2009). *Discovering Statistics using SPSS*, Sage Publications, London.

Friedrich S. et al. (2017). GFD: An R Package for the Analysis of General Factorial Designs, Journal of Statistical Software, to be published.

Gao, X. and Alvo, M. (2005). A nonparametric test for interaction in two-way layouts. *Canadian Journal of Statistics*, Volume 33, Issue 4, pp 529–543.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research,* 42(3), 237-288.

Hahn, S., Konietschke, F., Salmaso, L. (2014). *A comparison of efficient permutation tests for unbalanced ANOVA in two by two designs and their behavior under heteroscedasticity* Topics in Statistical Simulation, Springer Proceedings in Mathematics & Statistics Volume 114, pp 257-269.

Hájek, J. (1969). *A Course in Nonparametric Statistics*. Holden-Day, San Francisco.

Hájek, J., Šidák, Z., Sen, P.K. (1999). *Theory of Rank Tests*, Academic Press, Kent U.K.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. New York, Wiley.

Hettmansperger, T.P. , McKean, J.W. (2011). *Robust Nonparametric Statistical Methods*. CRC Press, Boca Raton.

Higgins, J.J., Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World 1,* 1994, pp 201-211.

Hodges, J.L. and Lehmann, E.I. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics* 33, pp 482-497.

Hora, S.C., Conover, W.J. (1984). The F Statistic in the Two-Way Layout with Rank-Score Transformed Data, *Journal of the American Statistical Association*, Vol. 79, No. 387, pp. 668-673.

Huang, M.L. (2007). A Quantile-Score Test for Experimental Design. *Applied Mathematical Sciences*, Vol. 1, No 11, pp 507-516.

Ito, P.K. (1980): *Robustness of Anova and Manova Test Procedures.* Handbook of Statistics, Vol. 1, (P.R.Krishnaiah, ed.), pp 199-236

Kaptein, M., Nass, C., Markopoulos, P. (2010). *Powerful and Consistent Analysis of Likert-Type Rating Scales*. CHI 2010: 1001 Users, April 10–15, 2010, Atlanta, GA.

Kay M. and Wobbrock J. (2015). *ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs.* R package version 0.9.5, URL: https://github.com/mjskay/ARTool .

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995). Robust and powerful nonorthogonal analyses. *Psychometrika*, 60, 395-418.

Konar, N.M. , Dag, O. and Dolgun, N.A.B. (2015). *Effects of Non-normality and Heterogeneity on Tests for One-Way Independent Groups Design: Type I Error and Power Comparisons.* CEB-EIB conference 2015, Bilbao.

Lachenbruch, P.A. and Clements, P.J. (1991). Anova, Kruskal-Wallis, Normal Scores and Unequal Variances. Communications in Statistics - Theory and Methods, 20 (1), pp 107-126

Leys, C., Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*, 46, pp 684-688.

Lei, X., Holt, J.K., Beasley,T.M. (2004). Aligned Rank Tests As Robust Alternatives For Testing Interactions In Multiple Group Repeated Measures Designs With Heterogeneous Covariances. *Journal of Modern Applied Statistical Methods*, 2004, Vol 3, Issue 2, pp 462-475.

Li Qinglong 2015: *Statistical Methods for Ranking Data, R Package StatMethRank.* URL: https://cran.r-project.org/web/packages/StatMethRank/StatMethRank.pdf .

Lindman, H. R. (1974): *Analysis of variance in complex experimental designs*. San Francisco: W. H. Freeman & Co.

Lix L.M., Keselman J.C. and Keselman, H.J. (1996). Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Review of Educational Research*, Vol. 66, No. 4, pp. 579-619.

Luepsen, H (2014). *R Functions for the Analysis of Variance*. URL: http://www.uni-koeln.de/~luepsen/R/ .

Luepsen, H. (2015). *Varianzanalysen - Prüfung der Voraussetzungen und Übersicht der nichtparametrischen Methoden sowie praktische Anwendungen mit R und SPSS.* URL: http://www.uni-koeln.de/~luepsen/statistik/texte/nonpar-anova.pdf URL: http://kups.ub.uni-koeln.de/6851/1/nonpar-anova.pdf .

Luepsen, H (2016). *The Lognormal Distribution and Nonparametric Anovas - a Dangerous Alliance.* URL: http://www.uni-koeln.de/~luepsen/statistik/texte/lognormal-anova.pdf

Luepsen, H (2017). The Aligned Rank Transform and discrete Variables - a Warning. in press, to appear in *Communications in Statistics - Simulation and Computation*

Mansouri, H. , Chang, G.-H. (1995). A comparative study of some rank tests for interaction . *Computational Statistics & Data Analysis* 19 (1995) 85-96 .

Mansouri, H. (1999a). Aligned rank transform tests in linear models. *Journal of Statistical Planning and Inference*, 79, pp 141-155

Mansouri, H. (1999b). Multifactor analysis of variance based on the aligned rank transform technique. *Computational Statistics & Data Analysis*, 29, pp 177-189

Mansouri, H. , Paige, R., Surles, J. G. (2004). Aligned rank transform techniques for analysis of variance and multiple comparisons. Missouri University of Science and Technology *Communications in Statistics - Theory and Methods* - Volume 33, Issue 9.

McSweeney, M. (1967). *An empirical study of two proposed nonparametric tests for main effects and interaction* (Doctoral dissertation, University of California-Berkeley, 1968). Dissertation Abstracts International, 28(10), 4005.

Osborne, J.W. (2008). *Best Practices in Quantitative Methods*. Sage Publications.

Puri, M.L. & Sen, P.K. (1985). *Nonparametric Methods in General Linear Models*. Wiley, New York.

Patrick, J.D. (2007). *Simulations to analyze Type I Error and Power in the Anova F Test and nonparametric Alternatives.* Thesis, University of West Florida.

Peterson, K. (2002). Six Modifications Of The Aligned Rank Transform Test For Interaction. *Journal Of Modem Applied Statistical Methods*. Vol. 1, No. 1, pp 100-109.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/ .

Richter, S.J. and Payton, M. (1999). Nearly exact tests in factorial experiments using the aligned rank transform. *Journal of Applied Statistics,* Volume 26, Issue 2.

Richter, S.J. and Payton, M. (2003a). *An Improvement to the Aligned Rank Statistic for Two-Factor Analysis of Variance.* Joint Statistical Meeting of the American Statistical Association, Journal of Applied Statistical Science, 14(3/4), pp 225-236.

Richter, S. J. and Payton, M. E. (2003b). Performing Two Way Analysis of Variance Under Variance Heterogeneity, *Journal of Modern Applied Statistical Methods*, 2 (1), pp 152-160.

Salazar-Alvarez, M.I. , Tercero-Gomez, V.G., Temblador-Pérez, M., Cordero-Franco, A.E., Conover, W.J. (2014). *Nonparametric analysis of interactions: a review and gap analysis*. Proceedings of the 2014 Industrial and Systems Engineering Research Conference, Y. Guan and H. Liao (eds.).

Salter, K.C. and Fawcett, R.F. (1993). The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation* 22 (1), pp 137-153.

SAS Institute, Inc. (2009). *SAS Users's Guide: Statistics*. Cary , N.C., SAS Institute.

Sawilowsky, S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research* 60, pp 91–126.

Scheirer, J., Ray, W.J., Hare, N. (1976). The Analysis of Ranked Data Derived from Completely Randomized Factorial Designs. *Biometrics*. 32(2). International Biometric Society, pp 429−434.

Shah, D. A., Madden, L. V. (2004). Nonparametric Analysis of Ordinal Data in Designed Factorial Experiments . *The American Phytopathological Society,* Vol. 94, No. 1, pp 33 - 43.

Sheskin, D.J. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall.

Shirley, E.A. (1981). A distribution-free method for analysis of covariance based on ranked data. *Journal of Applied Statistics*, 30 (2), pp 158-162.

IBM SPSS (2012). *IBM SPSS Statistics User's Guide*. Chicago, IBM Corporation.

Terpstra, J.F., McKean, J.W. (2005). Rank-Based Analyses of Linear Models Using R. *Journal of Statistical Software,* Volume 14, Issue 7, pp 1-26.

Thomas, J.R., Nelson, J.K. and Thomas, T.T. (1999). A Generalized Rank-Order Method for Nonparametric Analysis of Data from Exercise Science: A Tutorial. *Research Quarterly for Exercise and Sport, Physical Education, Recreation and Dance*, Vol. 70, No. 1, pp 11-23.

TIBCO Spotfire S+ 8.2 User's Guide 2010). TIBCO Software Inc.

Tomarken, A.J. and Serlin, R.C. (1986). Comparison of ANOVA Alternatives Under Variance Heterogeneity and Specific Noncentral Structures. *Psychological Bulletin*, Vol. 99, No 1, pp 90-99.

Toothaker, L.E. and De Newman (1994). Nonparametric Competitors to the Two-Way ANOVA. *Journal of Educational and Behavioral Statistics*, Vol. 19, No. 3, pp. 237-273.

Vallejo, G., Ato, M., Fernandez, M.P. (2010). A robust approach for analyzing unbalanced factorial designs with fixed levels. *Behavior Research Methods*, 42 (2), 607-617

van der Waerden, B.L. (1953). *Order tests for the two-sample problem. II, III*, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Serie A, 564, pp 303–310 and pp 311–316.

Wilcox, R. R. (1995). ANOVA: The practical important of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology,* 48, pp 99-114.

Wilcox, R.R. (2005). *Introduction to robust estimation and hypothesis testing*. Burlington MA, Elsevier.

Winer, B.J. et al. (1991): *Statistical Principles in Experimental Design*, McGraw-Hill

Wobbrock, J. O., Findlater, L., Gergle, D. & Higgins, J. (2011). *The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures*. Computer Human Interaction - CHI , pp. 143-146.

Yates, H.L. (2008). *A Comparison of Type I Error and Power of the Aligned Rank Method using Means and Medians for Alignment.* Emporia State University, Report for the degree of master of science.

Zimmerman, D.W. (1998). Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation ofTwo Assumptions. *The Journal of Experimental Education*, Vol. 67, No. 1 (Fall, 1998), pp. 55-68.

Zimmerman, D.W. (2004). Inflation of Type I Error Rates by Unequal Variances Associated with Parametric, Nonparametric, and Rank-Transformation Tests. *Psicológica*, 25, pp 103-133.

11. Appendix: Computational procedure for the ATS statistic

A more detailed description can be found in Brunner & Munzel (2002, chapter 3) as well as Akritas, Arnold and Brunner (1997).

First some notations for matrices shall be defined. For an integer n:

$\boldsymbol{I}_n$ is the identitiy matrix of order n,
$\boldsymbol{H}_n$ is a quadratic matrix of order n consisting of only 1 and
$\boldsymbol{P}_n = \boldsymbol{I}_n - (1/n)\boldsymbol{H}_n$

Based on the ranks R_{ijk} of the observed values y_{ijk} the following means

$$\bar{R}_{ij\cdot} = \frac{1}{n_{ij}}\sum_k^{n_{ij}} R_{ijk} \qquad \bar{R}_{i\cdot\cdot} = \frac{1}{J}\sum_j^{J} \bar{R}_{ij\cdot} \qquad \bar{R}_{\cdot j\cdot} = \frac{1}{I}\sum_i^{I} \bar{R}_{ij\cdot} \qquad \bar{R} = \frac{1}{I}\frac{1}{J}\sum_i^{I}\sum_j^{J} \bar{R}_{ij\cdot}$$

and cell variances

$$s_{ij}^2 = \frac{1}{N^2(n_{ij}-1)}\sum_k^{n_{ij}} (R_{ijk} - \bar{R}_{ij\cdot})^2$$

are computed. Let $\mathbf{V}_N$ be the diagonal matrix

$$V_N = N \cdot diag\left(\frac{s_{11}^2}{n_{11}}, \ldots, \frac{s_{IJ}^2}{n_{IJ}}\right)$$

and

$$S_0^2 = \sum_i^I \sum_j^J \sum_k^{n_{ij}} (R_{ijk} - \bar{R}_{ij\cdot})^2 / (n_{ij}(n_{ij}-1))$$

then the statistic for testing the main effect A

$$F_A = \frac{IJ^2}{(I-1)S_0^2}\sum_i^I (\bar{R}_{i\cdot\cdot} - \bar{R})^2$$

is approximately F distributed with degrees of freedom (f_A, f_0) with

$$f_A = (I-1)^2 S_0^4 / [(IJN)^2 tr(T_A V_N T_A V_N)]$$

$$f_0 = S_0^4 / \left[N^4 \sum_i^I \sum_j^J \left(\frac{s_{ij}^2}{n_{ij}}\right)^2 / (n_{ij}-1)\right]$$

where $T_A = P_I \otimes \frac{1}{J} H_J$ with the Kronecker product $\otimes$ and H as defined above. Similarly the test of main effect B is performed. Finally the statistic for testing the interaction effect AB:

$$F_{AB} = \frac{IJ}{(I-1)(J-1)S_0^2}\sum_i^I \sum_j^J (\bar{R}_{ij\cdot} - \bar{R}_{i\cdot\cdot} - \bar{R}_{\cdot j\cdot} + \bar{R})^2$$

with is also approximately F distributed with degrees of freedom (f_{AB}, f_0), where

$$f_{AB} = (I-1)^2 (J-1)^2 S_0^4 / [(IJN)^2 tr(T_{AB} V_N T_{AB} V_N)]$$

with $T_{AB} = P_I \otimes P_J$ and f_0 as above.