The Lognormal Distribution and Nonparametric Anovas - a Dangerous Alliance

Version 1
(4.4.2016)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Kontakt: Luepsen@Uni-Koeln.de

Universität zu Köln

The Lognormal Distribution and Nonparametric Anovas - a Dangerous Alliance

Abstract

Results from several simulation studies in the last two decades showed that ranking procedures for the comparison of means may lead to an inflation of the type I error rate when the underlying distribution is lognormal and the variances are heterogeneous, even in the case of equal sample sizes. In this study the error rates of the parametric F-test as well as those of seven nonparametric tests are compared in a two-way between subjects anova design. The methods under consideration are: 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 for the tests of the null model are computed for several lognormal distributions with varying degrees if skewness, varying sample sizes from 5 to 50, several degrees of variance heterogeneity as well as for balanced and unbalanced designs. It is shown that the error rates of main and interaction effects for all nonparametric methods increase above any acceptable limit for moderate cell counts of 20 and more, while the parametric F-test keeps the error completely under control at least in the case of equal sample sizes. These results show that nonparametric methods are not always acceptable substitutes for parametric methods such as the F test in research studies when parametric assumptions are not satisfied.

1. Introduction

The lognormal distribution is very common in practice. Typically the blood pressure (diastolic and systolic), the income and the consumption, e.g. of alcohol, are lognormally distributed. Characteristics: there is an absolute zero-point - by means of a suitable shift of the variable this must not be zero in practice - and a long tail on the right. When such data have to be analyzed introductory textbooks usually recommend applying nonparametric procedures because such methods are believed to be superior to normal theory techniques if data are nonnormal distributed (see e.g. Zimmerman & Zumbo, 1993). „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 analysis of variance (aov) with the assumptions of normality and variance homogeneity.

Some of the advocates should be mentioned at this point. Sawilowsky (1990) 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. Higgins & Tashtoush (1994) as well as Salter & Fawcett (1993) showed that the ART procedure is valid concerning the type I error rate and that it is 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 & Chang (1995) showed that the ART performs better than the F-test concerning the power in various situations with skewed and tailed distributions. Sheskin (2004) reported that the van der Waerden-test in the 1-factorial version beats the classical aov in the case of violations of the assumptions. Danbaba (2009) compared for a simple 3*3 two-way design 25 rank tests with the parametric F-test. 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.

Looking concretely onto the lognormal distribution a simulation study by Danbaba (2009) has to be mentioned: He showed that for several rank tests (including the RT, ART, INT and ATS) the error rates stay predominantly in the acceptable range and tend to decrease for increasing cell counts. Unfortunately he considered not the case of heterogeneous variances.

But on the other side inflated type I error rates have been revealed if ranking procedures are applied on skewed data with unequal variances, even in the case of equal cell counts. Zimmerman (1998) concurrently violated in a simulation study the assumptions of parametric tests (normality and homogeneity of variance) for various combinations of nonnormal distribution shapes and degrees of variance heterogeneity. He found that the type I error probability of the nonparametric Wilcoxon-Mann-Whitney rank-test (U-test) to be biased to a far greater extent than that of its parametric counterpart, the Student t-test. In a later study Zimmerman (2004) compared again the same tests for 25 distributions. He had been confronted with error rates rising to 40 percent for the U-test with n=20 and up to 75 percent with n=50 while the ratio of the two standard deviations increases to 3 in the case of an underlying lognormal distribution. And this even in the case of equal cell counts. Similar results were found for other skewed distributions although not to the same extent. There exist even earlier studies (e.g., Harwell, 1990; Rogan and Keselman, 1977) in which changes in the Type I error rates of the Student t-test for equal sample sizes are reported.

A more recent study comes from Carletti & Claustriaux (2005) who compared the ART-technique with the parametric F-test for the lognormal distribution and used a 2*4 design with a relation of 4 and 8 for the ratio of the largest to the smallest variance. They found that in the case of heteroscedasticity the ART has far 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. In addition the type I error increases up to 30 percent with larger cell counts. But they proposed an amelioration of the ART technique: to transform the ranks obtained from the ART according to the INT method, i.e. transforming them into normal scores (see chapter 2). This method leads to a reduction of the type I error rate, especially in the case of unequal variances.

At least, after numerous simulation studies and many theoretical investigations, it is now generally accepted that the t- and F- tests are robust under violation of homogeneity of variance, as long as sample sizes are equal. Nevertheless some exceptions have been found (Tomarken & Serlin, 1986). Unfortunately these results cannot be transferred to the nonparametric methods. For the case of heterogeneous variances there exist a couple of adequate tests, e.g. by Welch, Welch & James, Brown & Forsythe and by Brunner, Dette & Munk (see e.g. Tomarken & Serlin, 1986 and G. Vallejo et al., 2010) which work also with unbalanced designs. But they all have problems with skewed data (see e.g. G. Vallejo et al., 2010, Keselman et al., 1995 and Tomarken & Serlin, 1986). Nevertheess there exist some modifications for the tests mentioned above, mainly based on robust estimators for means and variances (see e.g. Cribbie et al., 2010 and G. Vallejo et al., 2010).

In this study the type I error rates of 7 nonparametric aov methods, all based on ranking, as well as of the parametric F-test are compared for several lognormal distributions with varying shape parameters, for balanced and unbalanced designs and sample sizes n_i ranging from 5 to 50.

2. Methods to be compared

It follows a brief description of the methods compared in this paper.

The anova model shall be denoted by

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

with fixed effects α_i (factor A), β_j (factor B), $\alpha\beta_{ij}$ (interaction AB) and error e_{ijk} .

RT (rank transform)

The rank transform method (RT) is just transforming the dependent variable (dv) into ranks and then applying the parametric aov 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. 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.

INT (inverse normal transform)

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

$$\Phi^{-1}(R_i/(n+1))$$

where R_i are the ranks of the dv 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 et al. (2009) found out that also the INT procedure results in slightly too high type I error rates if there are significant main effects.

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 aov. 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 aov to them. This procedure dates back to Hodges & Lehmann (1962) and had been made popular by Higgins & Tashtoush (1994) who extended it to factorial designs. In the simple 2-factorial case the alignment is computed as

$$x'_{ijk} = e_{ijk} + (\alpha\beta_{ij} - \alpha_i - \beta_j + 2\mu)$$

where e_{ijk} are the residuals and α_i, β_j, $\alpha\beta_{ij}$, μ are the effects and the grand mean. 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.

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.

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 the dv, conducting a parametric aov 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 are

$$\chi^2 = \frac{SS_{effect}}{MS_{total}}$$

where SS_{effect} is the sum of squares of the considered effect and MS_{total} is the total mean square. 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 is not diminished by other factors. 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 (see Wikipedia and van der Waerden (1953)) is an alternative to the 1-factorial aov by Kruskal-Wallis. The procedure is based on the INT transformation (see above). But instead of using the F-tests from the parametric aov, 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. Marascuilo and McSweeney (1977) transferred it to the case of repeated measurements. Sheskin (2004) reported that this procedure in the 1-factorial version beats the classical aov in the case of violations of the assumptions. 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 aov. Based on the relative effect (see Brunner & Munzel (2002)) the authors developed two tests to compare samples by means of comparing these relative effects: ATS (anova type statistic) and 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 the dv this method is suitable also for variables of ordinal or dichotomous scale. The rather complicated procedure is described by Akritas, Arnold and Brunner (1997) as well as by Brunner & Munzel (2002).

3. The Study

This is a 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 for the various aov methods.

In the current study only grouping (between subjects) factors A and B are considered. It examins 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_i per cell (total n=100) and a ratio $\max(n_i)/\min(n_i)$ 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. (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.)

As from the results of the studies by Zimmerman as well as those by Carletti & Claustriaux mentioned above to be expected, preliminary tests for the lognormal distribution revealed: all nonparametric methods under consideration here show rising type I error rates for increasing sample sizes in the case of heterogeneous variances. For a more precise investigation the error rates of the lognormal distribution had been computed for several parameters in the range n=5,10,...,50: varying

- the variance ratio (equal variances, slightly unequal variances (factor 2) and strongly unequal variances (factor 4)),
- variance heterogeneity on one factor (B) as well as on both factors (A and B),
- the skewness by means of the parameter σ (slightly skewed (σ=0.25), medium skewed (σ=0.50) and strongly skewed (σ=1.0)), and
- the variance by means of the parameter μ (small variance (μ=0), medium variance (μ=1) and large variance (μ=2), only for the parameter σ=0.50.

Additionally two lognormal distributions had been studied which both could be used for modelling the diastolic blood pressure dbp ($\bar{x} \sim 75$ and s~10 assumed), with

- parameters μ=3.5 and σ=0.3 resulting in $\bar{x} \sim 35$ (model I) and
- parameters μ=4.3 and σ=0.13 resulting in $\bar{x} \sim 75$ (model II).

Both have a standard deviation s~10. If an absolute minimum of 40 is assumed for dbp, model I gives a good fit for the difference dbp-40, while model II fits dbp untransformed. Both distributions have nearly the same shape (see figure 1 where for comparison purposes the density of the corresponding normal distribution is also plotted).

In the cases of variance heterogeneity the parameters μ and σ had been modified for some groups respectively cells to achieve the desired variance ratio. The parameters had been chosen so that the groups respectively cells compared had always equal cell means, see table 1. The cells modified depend on the design. Subsequently i,j refer to the indices of factors A respectively B.

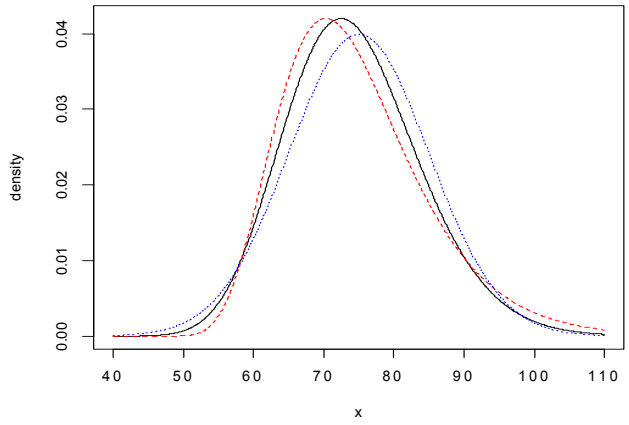


Figure 1: densities of lognormal distributions with μ=4.3 and σ=0.13 (solid black), μ=3.5 and σ=0.15 (dashed red) and normal distribution with μ=75 and σ=10 (blue points)

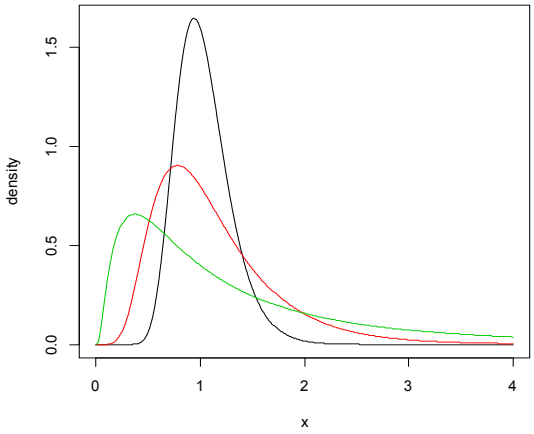


Figure 2: densities of several lognormal distributions: with parameters μ=0 and σ=0.25 (black), μ=0 and σ=0.5 (red), μ=0 and σ=1 (green)

- For the 2*4 balanced as well as the 4*5 unbalanced design and unequal variances on B the cells with $j \leq 2$ have a variance ratio of 4 and those with j=3 a ratio of 2.
- In the case of the 2*4 balanced design 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 $j \leq 3$ a ratio of 2.
- In the case of the 4*5 unbalanced design and unequal variances on A and B the cells with $i \leq 2$ and $j \leq 2$ have a variance ratio of 4 and those with $i \geq 3$ and $j \leq 3$ a ratio of 2.

If there are only slightly unequal variances, all cells listed above have a variance ratio of only 2.

Special attention is paid to heterogeneous variances in conjunction with unequal cell counts. As it is well known meanwhile, the F-test behaves conservative if large variances coincide with larger cell counts (*positive pairing*) and that it behaves liberal if large variances coincide with smaller cell counts (*negative pairing*) (see e.g. Feir & Toothaker, 1974 and Weihua Fan, 2006). Therefore the pattern of the s_i^2 had been chosen so that n_i and s_i^2 are independent.

The main focus had been laid upon the control of the type I error rates for α=0.05 for the various methods and situations. Therefore the error rates had been checked for both main effects as well as the interaction effect for the case of the null model (equal means).

4. Results

All tables are available online under the denotation *Appendix 6*:
http://www.uni-koeln.de/~luepsen/statistik/texte/comparison-tables/
Each table includes the results for all 8 methods and reports type I error rates as the proportions of rejections of the corresponding null hypothesis at α=0.05: for all 3 effects (factor A, factor B and the interaction), for n_i = 5,..,50 with equal and unequal cell frequencies, and for the lognormal distributions with various parameters. The tables are referred here as A *n.n.n*.

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 a deviation of 25 percent ($\alpha + 0.25\alpha$) - that is 6.25 percent for α=0.05 - is 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.

variances	μ	σ	$\bar{x}$	s^2	skewness
slightly skewed					
equal	0	0.25	1.0317	0.0686	0.778
unequal - ratio 2	-0.0295	0.3483	1.0317	0.1373	1.124
unequal - ratio 4	-0.0836	0.4791	1.0317	0.2746	1.655
medium skewed - small variances					
equal	0	0.5	1.1331	0.3650	1.750
unequal - ratio 2	-0.1001	0.6709	1.1331	0.7300	2.691
unequal - ratio 4	-0.2546	0.8712	1.1331	1.4588	4.409
strongly skewed					
equal	0	1	1.6487	4.6701	6.185
unequal - ratio 2	-0.2342	1.2167	1.6487	9.3400	11.781
unequal - ratio 4	-0.5317	1.4365	1.6487	18.683	25.887
medium skewed - medium variances					
equal	1	0.5	3.0802	2.6948	1.750
unequal - ratio 2	0.900	0.671	3.0802	5.3697	2.711
unequal - ratio 4	0.745	0.872	3.0802	10.8116	3.409
medium skewed - large variances					
equal	2	0.5	8.3729	19.9117	1.750
unequal - ratio 2	1.900	0.671	8.3729	39.6772	2.711
unequal - ratio 4	1.746	0.872	8.3729	79.7268	3.409
model I for diastolic blood pressure					
equal	3.5	0.3	34.640	10.63	1.963
unequal - ratio 2	3.459	0.415	34.640	21.26	2.108
unequal - ratio 4	3.385	0.565	34.640	42.52	2.402
model II for diastolic blood pressure					
equal	4.3	0.13	55.216	9.703	1.846
unequal - ratio 2	4.292	0.183	55.216	19.40	1.871
unequal - ratio 4	4.275	0.257	55.216	38.80	1.924

Table 1: Scheme of parameters for seven different types of lognormal distributions used in this study

equal variances

For nearly all different types of the lognormal distribution the type I error rates for all methods stay in the interval of moderate robustness, mostly even in the interval of stringent robustness. Only for the strongly skewed distribution with a skewness of 6.2 (parameters 0/1) the ART-method shows rates beyond the acceptable range, usually between 8 and 11 percent, especially for the main effects (see tables A 6.3.1 to 6.3.4). As a consequence also the values of the ART+INT-method are increased, but below 6 percent for cell counts $n_i \leq 30$. But also the parametric F-test reveals in this situation inflated rates (between 6 and 8): for the test of the interaction, although they decrease to 6 percent with rising n_i (see tables A 6.3.5 and 6.3.6).

unequal variances on B

Here again the strongly skewed distribution leads to a different behaviour of the ART- and the ART+INT-procedures. Whereas usually the rates for the tests of the main effect A and the interaction remain unaffected by the heterogeneity of factor B, they are highly raised for these two methods in the case of a strongly skewed distribution. But the values, usually clearly above 10 percent, tend to fall for increasing n_i . Here also the application of INT to the ART-technique shows a dampening effect (see table A 6.3.1 and A 6.3.2). The behaviour of the parametric F-test is the same as above in the case of homogeneity.

As to be expected the results are completely different for the test of factor B. The error rates of all nonparametric methods rise with increasing cell counts n_i , even if the variance ratio is only two. The extent differs a bit from the distribution parameters, especially from the skewness, and from the degree of variance heterogeneity. For the case of slightly heterogeneous variances the values rise to 10 for moderately skewed distributions (see e.g. A 6.1.3), up to 20 for medium skewed ones and up to 27 for strongly skewed distributions (see e.g. A 6.3.3 and figure 7). And for unbalanced designs the values lie even higher. For the case of strongly heterogeneous variances the rates rise generally up to 60 percent and more. And only for distribution model II, the one with the smallest skewness, all methods show acceptable error rates at least for small cell sizes $n_i \leq 20$ (see A 6.7.3). And this applies even in a larger extent to the ART- and the ART+INT-methods.

unequal variances on A and B

Also for heterogeneous variances on both factors the results depend on the degree of skewness as well as on the degree of variance heterogeneity. Only in the case of a small skewness the test of the interaction is not affected when the design is balanced (see A 6.1.5). Otherwise the error rates of the nonparametric tests rise for all tests to percentages between 7 and 70: for slightly heterogeneous variances up to 7 - 15 (small skewness) and 12 - 16 (large skewness), and for strongly heterogeneous variances up to 12 - 18 (small skewness), and 28 - 40 (large skewness). And for unbalanced designs the values lie even higher (see figure 8).

The behaviour of the parametric F-test is also affected, but independent of the cell counts n_i and independent of the degree skewness. For any variance heterogeneity this occurs in unbalanced designs where error rates between 6 and 8 are produced (see e.g. table A 6.1.2 and A 6.1.6 as well as figure 6). This confirms the „classical“ behaviour of the F-test.

As it appears in the specification of this study, there are, in the case of strong heterogeneity, cells with a variance ratio of 4 as well as a ratio of 2. There had been also simulations in which only a variance ratio of 4 had been used. Their results are partly different: for the test of factor A the rates were clearly larger in balanced designs and a bit smaller in unbalanced designs and for the

tests of factor B vice versa, whereas for the interaction effect the rates were generally larger. But these differences seem to depend on the specific pattern of those cells with larger variances. (The corresponding tables are not reproduced in appendix 6.)

differences between methods

As already mentioned the results for the ART are comparatively poor in the case of a strongly skewed lognormal distribution. The dampening effect of the normal transformation INT to the ART-results is only helpful for smaller n_i . Additionally the ART produces inflated rates in the case of unequal variances on B where all other methods seem to be unaffected (see e.g. A 6.1.1, 6.4.1, 6.7.1). On the other side the ART as well as the ART+INT perform better than the other nonparametric methods for the tests of the main effects in unbalanced designs when the variances are inhomogeneous on both factors (see e.g. A 6.1.2, 6.2.2, 6.5.2, 6.5.2, 6.7.2), though here also for larger n_i the rates rise beyond the acceptable limit.

For the other nonparametric methods there is only one remarkable result, though only of minor practical use: The Puri & Sen- and the ATS-method show acceptable error rates at least for smaller cell counts (Puri & Sen: $n_i \leq 15$ and ATS: $n_i \leq 25$) for the test of the interaction in unbalanced designs in the case of slightly unequal variances on both factors.

One final remark: for the parametric F-tests it was to be expected that the rates usually decrease with rising n_i according to the central limit theorem. But the same can be observed for the ART-technique in those cases where the results are not primarily affected by unequal variances, especially for the test of factor A in the case of heterogeneous variances on factor B (see e.g. A 6.1.2, 6.5.1, 6.5.2, 6.6.2) where the rates decrease from 6-7 to 4-5.

the parametric F-test

The parametric F-test is definitely the best performer concerning the type I error rate. In the cases of unequal variances it reacts slightly liberal in unbalanced designs. If the heterogeneity is restricted to only one factor (in this study factor B) the rates are almost acceptable for that factor (see figure 5). But if both have unequal variances especially the interaction is strongly affected with rates between 9 and 10 percent (see figure 6). Remarkable is the fact that the rates tend to fall with an increasing degree of skewness.

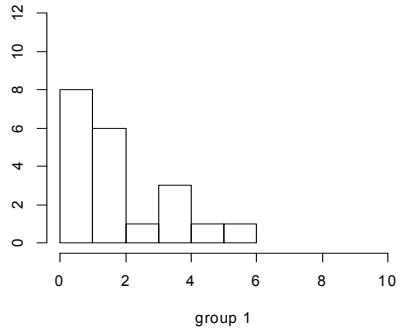
the models for the diastolic blood pressure

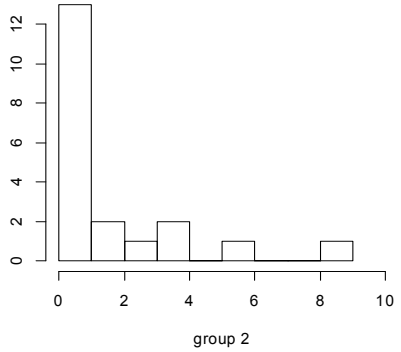
In chapter 3 two parameter sets for the lognormal distribution were presented to model the diastolic blood pressure. Looking at figure 1 there seems to be no difference between them. But the results for the error rates are slightly different, less concerning the quality than the quantity. Both models show inflated error rates for the same effects in the same situations. But for model I (parameters μ=3.5 and σ=0.3) the percentages are generally twice as high as for model II (parameters μ=4.3 and σ=0.13). For instance for the cases of small variance ratios, either on B or on both factors, the error rates are under control for small cell counts ($n_i \leq 15$) in model II (see figure 9), and for the case of large variance ratios on B the rates for the test of factor B rise up to 39 percent for model I but only up to 12 percent for model II (equal sample sizes). The reason might be the slight difference of skewness. Nevertheless the result is surprising: first that there are differences at all, having the two almost identical distribution shapes in mind, and secondly, as there are differences, that it is model II that fits better. Though it makes sense to subtract the minimum of 40 from the dbp before analyzing the variable in order to fit better the characteristics of the lognormal distribution.

5. An Explanation

Since the phenomenon of rising type I error rates occurs for all ranking based nonparametric methods the reason for that has to be searched in the ranking. First a simple example for a one-way anova with two groups on a lognormal distributed variable x shall illustrate the problem. The model for x is a lognormal distribution with parameters $\mu=0$ / $\sigma=1$ in group 1 and $\mu=-0.531$ / $\sigma=1.44$ in group 2 (see table 1 and figure 3). The values of x and its ranks are represented in table 2 together with the basic statistics and a histogram for both groups.

	group 1			group 2	
case no	x	rank(x)	case no	x	rank(x)
1	0.156	5	21	0.032	1
2	0.256	9	22	0.072	2
3	0.335	11	23	0.129	3
4	0.644	15	24	0.131	4
5	0.694	16	25	0.170	6
6	0.756	17	26	0.206	7
7	0.824	18	27	0.246	8
8	0.892	20	28	0.276	10
9	1.106	22	29	0.339	12
10	1.206	23	30	0.364	13
11	1.267	24	31	0.397	14
12	1.515	26	32	0.857	19
13	1.587	27	33	0.982	21
14	1.741	28	34	1.432	25
15	2.018	30	35	1.986	29
16	3.020	32	36	2.581	31
17	3.135	33	37	3.415	34
18	3.420	35	38	3.728	36
19	4.041	37	39	5.174	38
20	5.901	39	40	8.215	40





	group 1	group 2
mean(x)	1.73	1.54
variance(x)	2.19	4.61
skewness(x)	1.24	1.71
mean(rank(x))	23.35	17.65

Table 2: raw values of x, their ranks rank(x), histograms and basic statistics for both groups

The results for the anova: the p-value is 0.748 for x and 0.123 for rank(x) which is clearly smaller. At first sight it is not obvious that both distributions reproduce the same mean (see figure 3) because the distribution for group 2 seems to have a smaller mean. But - what cannot be seen in the figure - the large number of small values is equalized by a few numbers of very large values, due to the skewness. (In case of a symmetric distribution there should be a similar number of values at both ends.) But this is no more valid for the ranks. While the original values at the left end lie very close together, their ranks are equally spaced, i.e.drift apart. And at the right tail the few values move closer together due to the ranking. Therefore group 2 will always have a considerably smaller mean rank than group 1. This is illustrated by the example: Most of the small ranks belong to group 2 while the major part of large ranks belong to group 1.

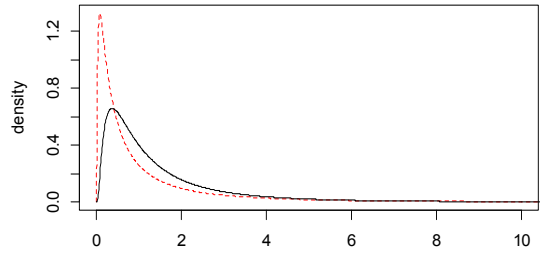


Figure 3: lognormal distributions with parameters μ=0 / σ=1 (solid black) for group 1 and μ=-0.531 / σ=1.44 (dashed red) for group 2, both reproducing the same mean $\bar{x}$ = 1.65.

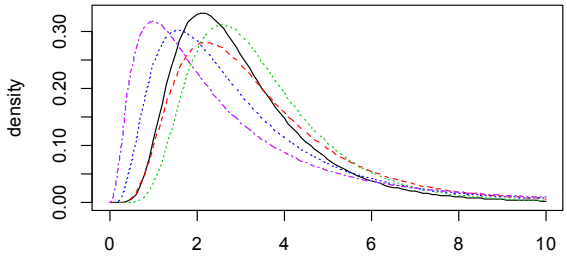


Figure 4: lognormal distribution (μ=1 / σ=0.5) (solid black) with several alternatives: different mean (green), different mean and variance (red), equal means and variance ratio 2 (blue) as well as variance ratio 4 (magenta).

6. Conclusion

The first conclusion is evident: As even for variance ratios of two - which has to be considered as common - the error rates of main and interaction effects for all nonparametric methods increase above any acceptable limit for moderate cell counts of 20 and more, these procedures have to be avoided for variables with right skewed distributions similar to a lognormal distribution. Furthermore the comparison of the two models for the diastolic blood pressure reveals the dependency of the tests on small variations of the distributional parameters. As a consequence: if a comparison of several groups yields a significant result, one cannot be sure that it is caused by unequal means. It could be unequal variances as well.

On the other side the parametric F-test keeps the error rate completely under control, at least for equal cell counts.

These results show that nonparametric methods are not always acceptable substitutes for parametric methods such as the F test in research studies when parametric assumptions are not satisfied.

But at the end a quite different question has to be put: Is it reasonable at all to treat means and variances separate, i.e. to compare means assuming equal variances? Perhaps it is more realistic for strongly right skewed data to assume a distribution for the alternative hypothesis which differs not only in regard to the mean but also to the variance. Figure 4 shows a lognormal distribution (parameters μ=1 / σ=0.5) with 4 alternatives. The shape of those two with equal means and unequal variances suggest rather the opposite: unequal means and equal variances. Therefore such a model does not seem to be realistic.

7. Software

This study has been programmed in R (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). All the computations had been performed on a Windows notebook.

8. Figures

The parametric F-test

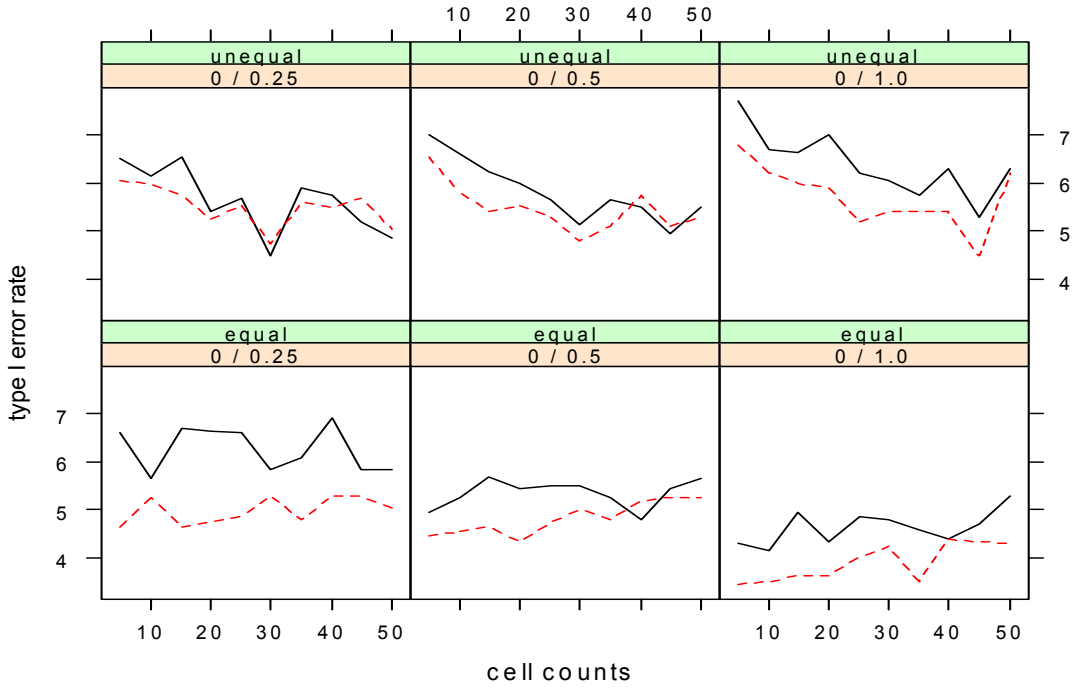


Figure 5: Type I error rates of the parametric F-test for the test of factor B with unequal variances: ratio 4 (solid black) and ratio 2 (dashed red), for equal/unequal cell counts and for 3 degrees of skewness.

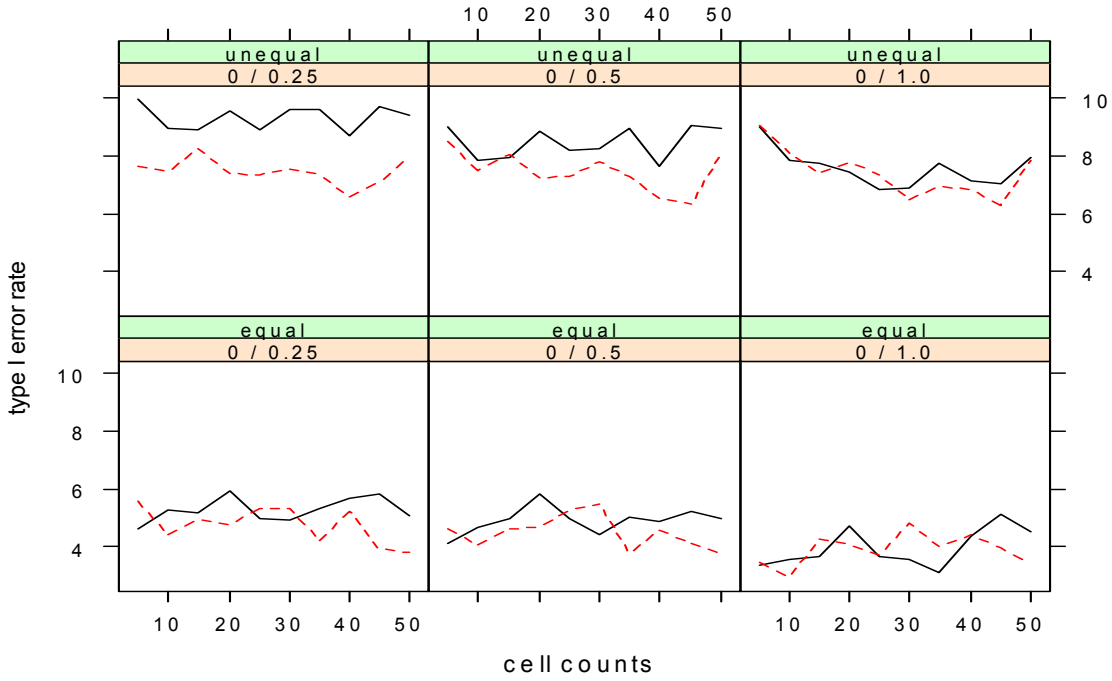


Figure 6: Type I error rates of the parametric F-test for the test of the interaction with unequal variances: ratio 4 (solid black) and ratio 2 (dashed red), for equal/unequal cell counts and for 3 degrees of skewness.

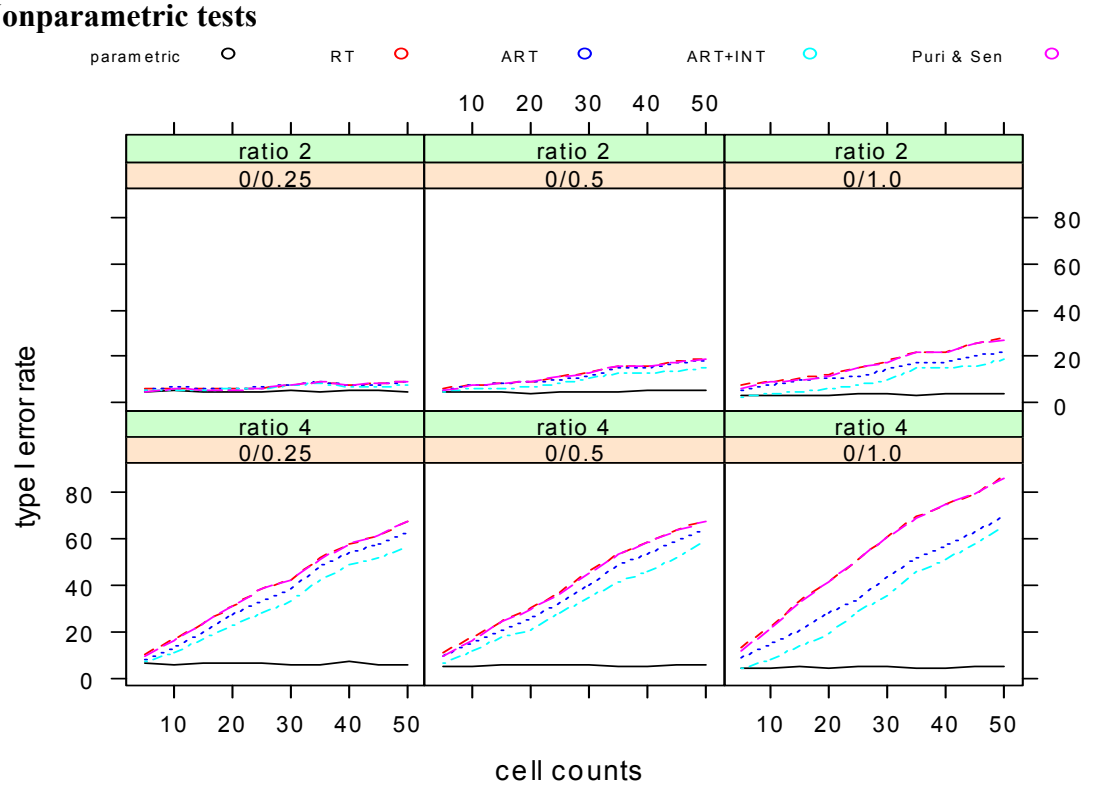


Figure 7: Type I error rate of several nonparametric methods for the test of factor B with 2 degrees of unequal variances (on factor B) and for 3 degrees of skewness (equal cell counts) .

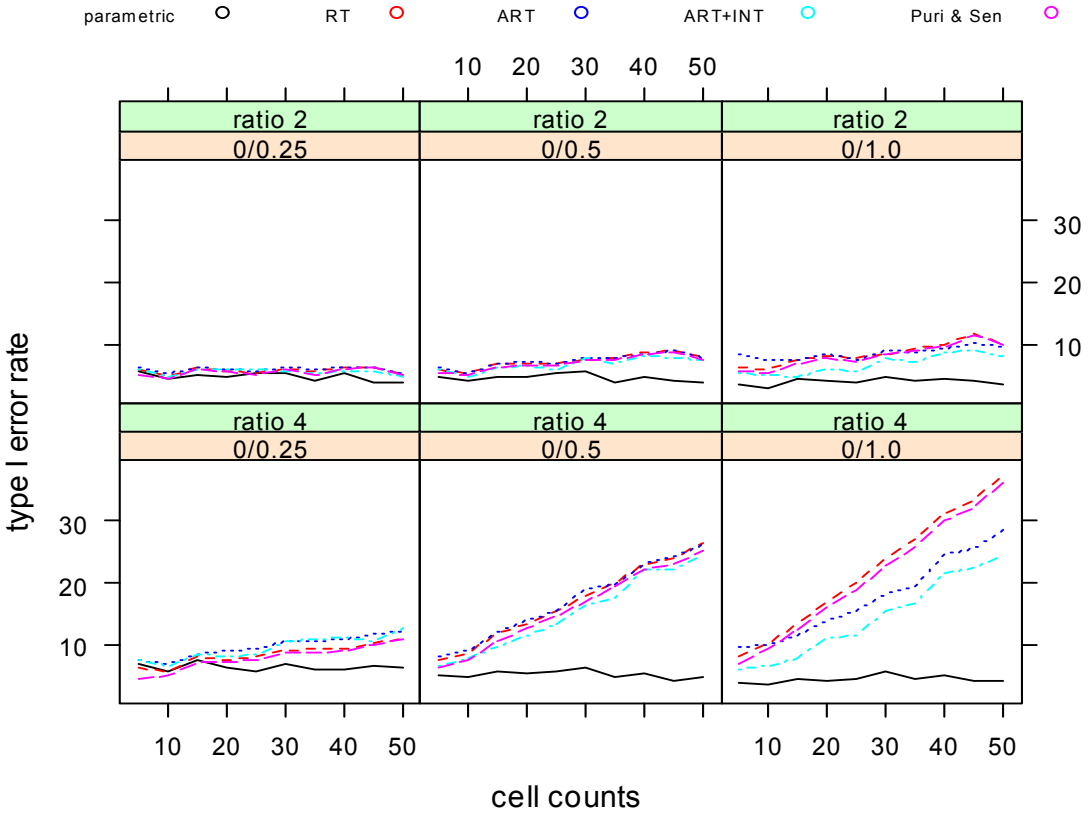


Figure 8: Type I error rate of several nonparametric methods for the test of the interaction with 2 degrees of unequal variances (on factors A and B) and for 3 degrees of skewness (equal cell counts) .

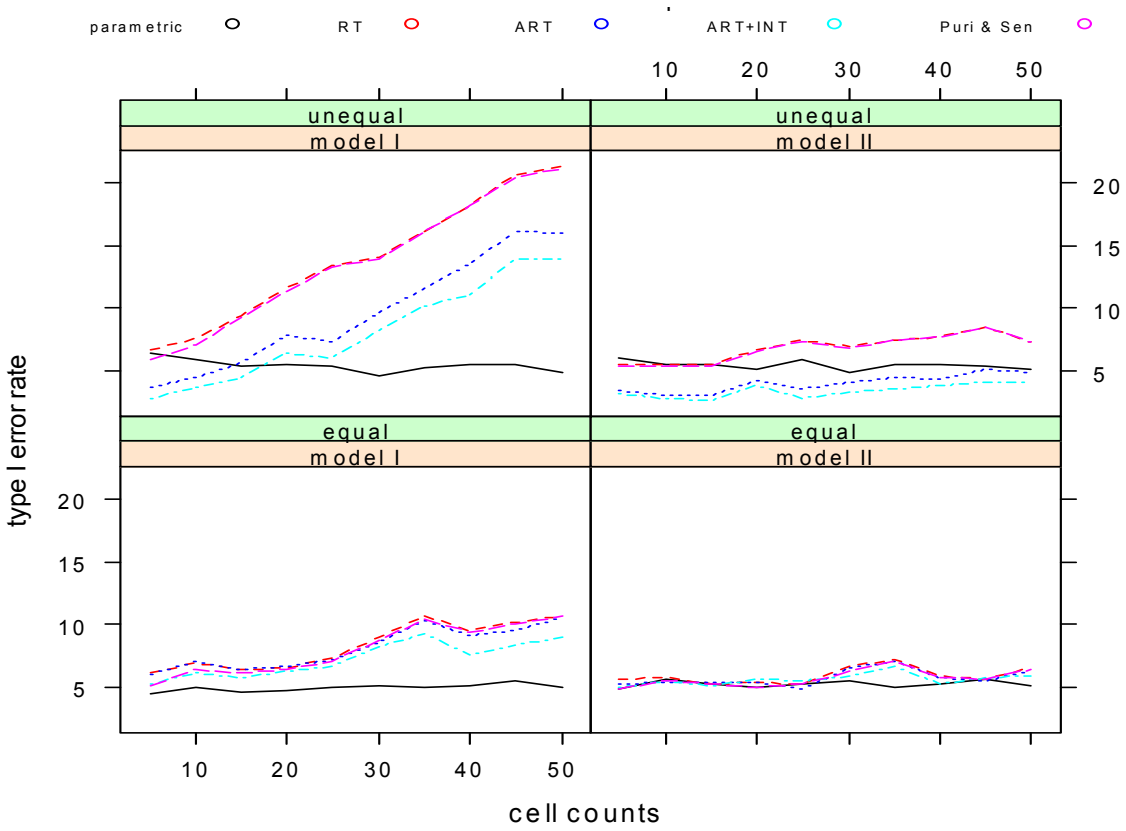


Figure 9: Type I error rate of nonparametric methods for the test of factor B with a small heterogeneity for the two dbp models as well as equal and unequal cell counts.

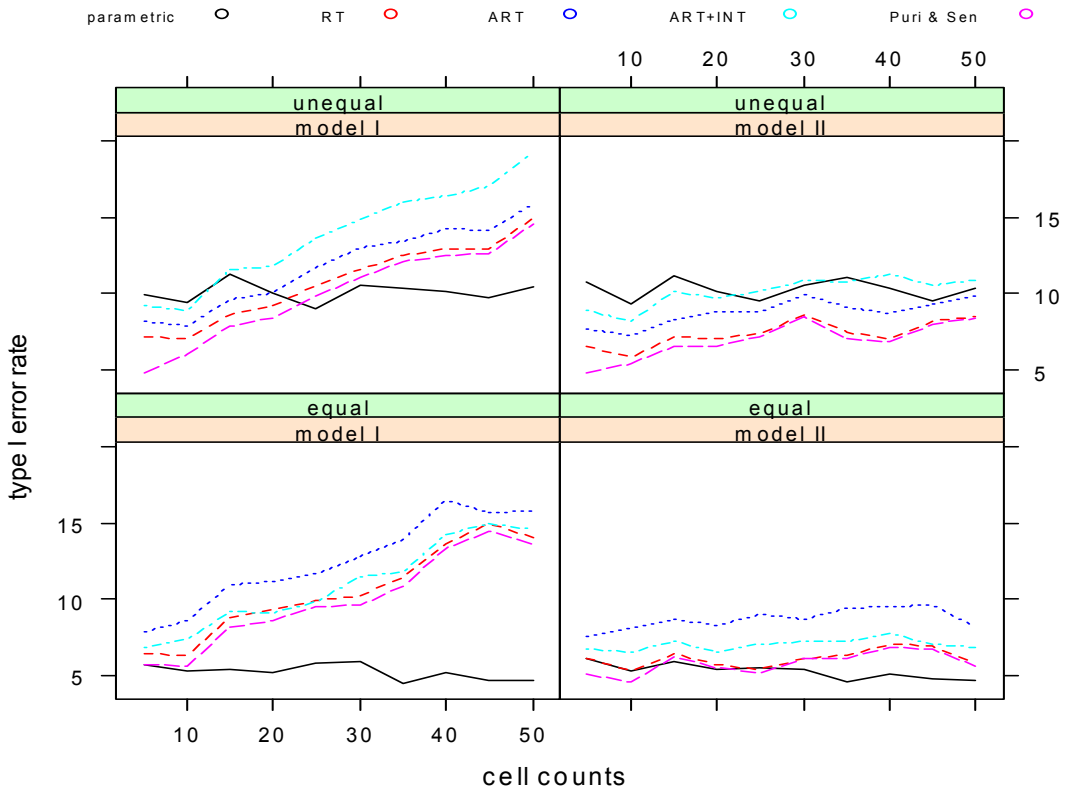


Figure 10: Type I error rate of nonparametric methods for the test of the interaction with a large heterogeneity on both factors for the two dbp models as well as equal and unequal cell counts.

9. 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? *Behavourial Genetics*, 39 (5), pp 380-395.

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

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

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

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.

Cribbie, R.A., Fikensenbaum, L., Keselman, H. J. & Wilcox, R.R. (2010). *Effects on Nonnormality on Test Statistics for One-Way Independent Groups Designs.* University of Manitoba (CA), http://home.cc.umanitoba.ca/~kesel/Cribbie_param_bootstrap_feb_2010.pdf

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

Fan, W. (2006). *Robust means modelling: An Alternative to Hypothesis Testing of Mean Equality in Between-subject Designs under Variance Heterogenity 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.

Harwell, M.R. (1990). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics*, 17, 297-313.

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* 27, pp 324-335.

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.

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

Luepsen, H (2014). *R-Funktionen zur Varianzanalyse*. URL: http://www.uni-koeln.de/~luepsen/R/ .

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. , 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.

Marascuilo, L.A., McSweeney, M. (1977): *Nonparametric and distribution- free methods for the social sciences.* Brooks/Cole Pub. Co.

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

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

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/ .

Rogan, J.C., & Keselman, H.J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal*, 14, 493-498.

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.

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: 158-162.

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.

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.

Wikipedia. URL: http://en.wikipedia.org/wiki/Van_der_Waerden_test .

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.

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., Zumbo, B.D. (1993). Rank Transformations and the Power of the Student t Test and Welch t' Test for Non-Normal Populations With Unequal Variances. *Canadian Journal of Experimental Psychology*, 1993, 47:3, pp. 523-539