Anova with binary variables - The F-test and some Alternatives

Version 3.0
(29.10.2019)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Kontakt: Luepsen@Uni-Koeln.de

Universität zu Köln

ANOVA with binary variables - The F-test and some Alternatives

Abstract

Several methods to perform an ANOVA with a binary dependent variable in 2-way layouts are compared with the parametric F-test. Equal and unequal cell counts as well as several different effect models are taken into account. Special attention has been paid to heterogeneous conditions, which are caused by nonnull effects through the relation of the binomial probability and its variance. For between subject designs Puri & Sen's L statistic, Brunner & Munzel's ATS, the χ^2-test of log-linear models, the logistic and the probit regresssions are considered. The L statistic is recommended, because the F-test cannot keep always the type I error under control, if there are nonnull effects. For mixed designs the Huynh-Feldt adjustment, Hotelling Lawley's multivariate test, Puri & Sen's L statistic, Brunner & Munzel's ATS, Koch's ANOVA, GLMM and GEE models are considered. None of these methods is able to cover all situations. Depending on the design and on the model to be checked, in most cases the parametric F-test with adjustment, the multivariate test or Koch's method are advised. Additional results: heterogeneous correlations and the size of the design have an impact, particularly on the F-test.

Keywords:

ANOVA, binary, dichotomous, Puri & Sen, ATS, GLMM, GEE, logistic, probit, regression.

1. Introduction

The analysis of variance (ANOVA) is one of the most important and frequently used methods of applied statistics, mainly for the analysis of designs with only grouping factors (between subject designs) and of designs with grouping and repeated measurements factors, usually referred as mixed or split-plot designs. (The term ANOVA is commonly used for the analysis of both designs, though the analysis of repeated measures designs may be subsumed under mixed models.) There is the parametric version and there are nonparametric methods as well. The first one has assumptions, of course. These are essentially normality of the residuals, homogeneity of the variances, and in the case of repeated measurements additionally sphericity and homogeneity of the covariance matrices over the groups. But what to do, if the dependent variable *y* is dichotomous, e.g. with values yes or no, or 1 and 0?

Due to the familiarity and simplicity of the ANOVA methodology, one could 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 error probability) are insensitive to departures from the assumptions on which it is derived.*" (See Ito, 1980). One of the first, who investigated the applicability of the parametric F-test to a dichotomous response variable, was Lunney (1970). His simulations showed that for 1-, 2- and 3-factorial designs the type I error rate is controlled as long as $N \geq 20$ for $0.2 \leq p \leq 0.8$, and $N \geq 40$ for other values of p (p being the percentage of one of the outcomes of y). And the power is satisfying as long as N is not too small, N being the df of the error term, approximately the total sample size. This seems reasonable as on one side the homogeneity of the variances is the most essential assumption - noting that the variance $p(1\text{-}p)$ depends on the mean p - and on the other side for $0.25 \leq p \leq 0.75$ variances of a binomial distributed outcome can be regarded as equal. Unfortunately Lunney's study has fundamental restrictions: first he examined only equal sample sizes, and secondly he checked type I error rates only if there are no other effects, and therefore neglecting the cases of unequal variances.

Besides that only between subject designs had been studied. D'Agostino (1971) wrote a detailed critic on Lunney's paper. Nevertheless it remains one of the most important works on this subject. Only decades later, Jaeger (2008) expressed concern on the use of the parametric ANOVA for the analysis of a binary outcome. Therefore alternatives are searched, for between subject as well as for split-plot designs.

From these the focus has been laid here on those methods, which are well-known and easily applicable in the software systems. First, one of the nonparametric ANOVA methods could be applied. See Luepsen (2017) for an overview for the case of between subject designs. To be considered here are the Puri & Sen-method (e.g. Puri & Sen, 1985), often referred as L statistic, the ANOVA-type statistic ATS (e.g. Brunner & Munzel, 2002) and a nonparametric ANOVA procedure proposed by Koch (1969). The L statistic and the ATS are available for both designs, whereas Koch's method is designed for mixed designs. They all are based on ranking the (usually continuous) observed values, but can be applied also to a binary outcome. In this case, the ranking, which is part of the algorithm, transforms the two values just into two other distinct values by using midranks for ties, thus having no real effect. The tests produce different test statistics, even if applied to binary variables. Other methods based on rank transformation, e.g. the rank transform (see Conover & Iman, 1981), or the inverse normal transformation (see e.g. Mansouri and Chang, 1995), would make no difference compared to the parametric F-test. The popular aligned rank transform (see e.g. Mansouri and Chang, 1995), is not reasonable for dichotomous variables, as Luepsen (2016) pointed out.

Additionally there are methods designed to analyze a dichotomous variable: log-linear models including the χ^2 test, logistic regression and probit regression for between subject designs (see e.g. Agresti, 2002), and the corresponding methods for dependent samples: GEE (*Generalized Estimating Equations*), established by Liang & Zeger (1986), and GLMM (*Generalized Linear Mixed Models,* sometimes also called MLM, *multi level models*) by Harville (1977). Both are extensions of the generalized linear models GLM allowing correlated responses. Finally, reflecting that unequal binomial probabilities p_i result in unequal variances, methods for mixed designs should be considered that do not assume sphericity, e.g. the Huynh-Feldt adjustment for the parametric ANOVA and the multivariate statistic by Hotelling-Lawley.

Of course, there exist a large number of studies concerning the methods listed above, but usually each compares only a couple of them. And the situations, which are investigated, differ from study to study. E.g. the designs or models are different, the sample sizes are varying, or often the type I error rates are controlled only for the null model. Just to mention the lack of models with nonnull interaction effects. Also the challenging pairing problem is rarely treated: the parametric F-test tends to be conservative, if cells with larger n_i have also larger variances (positive or direct pairing), and reacts liberal, if cells with larger n_i have the smaller variances (negative or inverse pairing), see e.g. Feir & Toothaker (1974). Therefore the results are inconsistent. So the aim of this research is to compare the most popular and in the literature most favored ANOVA methods for binary variables within a common frame of designs, models and situations, which should make them better comparable.

2. Literature Review

Although there are numerous studies comparing the different methods mentioned above, only few of them consider a binary response format. Therefore one has to look onto those, which include the impact of heterogeneous variances, because only for $.25 < p < .75$ the binomial distributions to be compared can be assumed to have equal variances. Furthermore it has to be accepted, that most of them deal only with 1-factorial designs and therefore give no information

about the behavior of the interaction effects. The results cited here arose from simulation studies. First the case of between subject designs.

Hsu & Feldt (1969) compared the 1-factorial ANOVA F-test with the χ^2 test, which may be considered as the obvious test for such a simple design, especially for the case of 2 values of *y*. First, the χ^2 test demands a minimum sample size that is higher than that required by the F-test. The second limitation of the χ^2 test is that it is not easily extended to factorial designs and tests of interactions. They confirm for this simple case the results found by Lunney (1970) and favor the F-test. One of the few studies considering log-linear models for the analysis of factorial designs comes from Swafford (1980). He explains in detail the problems arising when specific effects, especially interaction effects, have to be tested, because commonly hierarchic models are used, which do not guarantee independent tests of all effects. Tansey et al. (1996) compared log-linear models and logistic regression and list several advantages of the logistic approach in the case of ANOVA designs.

Concerning the L statistic by Puri & Sen one has to restrict mainly to studies of the Kruskal-Wallis test (KW), which is identical to the L statistic limited to one factor. Lix et al. (1996) reviewed articles dealing with the consequences of assumption violations for one-way ANOVAs, among them detailed studies by Tomarken & Serlin (1986) as well as Feir & Toothaker (1974), who analyzed the F-test and nonparametric alternatives under variance heterogeneity. They summarize that the KW appears to be sensitive to the presence of heterogeneous variances in both balanced and unbalanced designs, and that it is difficult to establish clear guidelines regarding the use of the KW under heterogeneity. Sawilowski (1990) reports several studies of factorial designs conducted by Harwell (e.g. Harwell et al. , 1992), which show the L statistic as a robust but conservative test, needing large samples (N>100) to achieve a reasonable power. To summarize: although Puri & Sen‘s method behaves rather conservative, it may be a good choice in cases of heterogeneity, which is a condition for its application on binary variables.

Logistic and probit regression seem to be ideal methods for analyzing a binary variable. Their disadvantage: the large *n* requirement. Malhotra (1983) compared them with OLS regression. In his simulation study he emphasized the effect of extreme p (0.1, 0.2, 0.8, 0.9). For smaller and medium sample sizes (<50) he sees the OLS regression superior to the logit and probit regressions, whereas for large samples (>100) he favors the logistic regression because of an up to 10% higher power rate. The relative performance of all three models is quite comparable at p=.5, regardless of sample size. Malhotra reported in his publication also quite a number of comparative studies and gave the results in a clearly arranged table. Nearly the same results were reported by Cleary & Angel (1984) and Pohlmann & Leitner (2003). In studies, considering both the logit and probit regressions, generally the logit approach is seen to be more efficient, but unfortunately none of the studies examined the tests together with nonnull interactions.

There are a number of studies related to the ATS. First to mention, Brunner et al. (1999a) who compared the ATS method with the KW in respect to unequal sample sizes, different pairing and unequal variances. They found the KW to react too liberal in the case of heterogeneous variances, even for equal n_i, whereas the ATS keeps the type I error completely under control. Unfortunately the ATS has type I error rates beyond the limit for small N. A comparable power has been observed for both tests. In the cases of positive and negative pairing, the ATS has its error rates closer to the α-level than the KW. Richter & Payton (1999) compared the ATS with the classical F-test in a 2-factorial study with heterogeneous variances, and state that the ATS keeps always the α level, but performs worse than the F test for small $n_i \leq 10$ regarding the power. It is virtually powerless to detect small to moderate effects, but getting nearer to the F-test

for increasing effect sizes. The general judgement: the ATS controls the type I error, except for small n_i, but possesses a poor power.

Among the first studies of applying the one-way parametric repeated measures ANOVA to a binary response, were those by Cochran (1950), Draper (1972) and Mandeville (1972). Cochran and Draper found in their simple simulations only neglectable violations of the type I error rate. Mandeville compared the F- and Q-test together with the multivariate statistic by Hotelling-Lawley for $.1<p<.9$, different correlations, but equal variances. He showed for the number of treatments $k > 5$ that the F-Test has generally the larger power and the lower type I error rate, at least for $N>60$, while the multivariate test reveals in some instances larger error rates. The test appears disappointing because, depending on the number of levels k and the correlations, it reacts sometimes rather liberal and sometimes too conservative, especially for extreme p: liberal mostly for small correlations r and conservative for large r. Stiger et al. (1998) evaluated the F-test with and without the Huynh-Feldt correction and the multivariate test for an ordinal 4-point scale in a split-plot design both with AR(1) covariance structure and $r=0.5$. For all three methods the error rates are rather close to the nominal level for both repeated measures effects, though the rates slightly increase, if the distribution of y is skewed. Generally the F-test without correction tends to be sometimes mildly liberal, while the Huynh-Feldt correction renders it more conservative. In regard to the power the multivariate appeared as the poorest. The author‘s recommendation: F-test with Huynh-Feldt correction, which seems to be the favorite also in other studies.

Concerning Puri & Sen‘s L statistic for within subject designs, there is only one study to mention: Harwell & Serlin (1994) compared F-test, Friedman test and L statistic in a one-way design with equal variances but nonspherical covariance matrices. When covariances are equal, all of the tests perform satisfactorily. For the 2:1 covariance ratio the L statistic performs well, while the F-test tends to produce inflated error rates for $k>3$. For covariance rations 3:1 and 5:1 the L statistic produces more and more inflated error rates, while the F-test performs poorly. In contrary to other findings Harwell & Serlin report that for nonnormal distributions the power of the L statistic was generally higher than the F test.

For the ATS in a mixed design there is a study by the authors themselves, Akritas & Brunner (1997b), in which they showed that the statistic keeps the α level correct, for equal and unequal covariance matrices. Konietschke et al. (2010) analyzed the ATS in a 1-factorial within subject design considering different covariance matrices and also a dichotomous dependent variable. They, too, attested the ATS a perfect control of the type I error.

Another solution for the analysis of split-plot designs is supplied by G. Koch, who proposed several nonparametric ANOVA procedures (Koch, 1969). There are a couple of comparisons taking Koch‘s method into consideration. Tandon & Moeschberger (1989) compared the F-test joined with the Huynh-Feldt correction, the multivariate approach and Koch‘s method in several mixed designs with different correlations r (0, 0.1, 0.25). In contrary to the parametric tests, Koch‘s test shows slightly liberal results for the group effect when $n_i \leq 10$. In contrast, the parametric tests offends the type I error rate for the tests of the repeated measures effect and the interaction, while Koch‘s method is more conservative. For the case of unequal correlations Koch‘s test performs the best, whereas the corrected F-test behaves conservative. One disadvantage is the poor power for small N. Ernst & Kepner (1993) come in their simulation study to similar results.

Meanwhile a large number of studies are concerned with GEE and GLMM, but only very few compare these methods with the parametric F-test. As both methods are based on large sample

asymptotic theory, it is not surprising that the tests of the parameters are generally liberal for small samples $N<50$ (see e.g. Qu et al., 1994 and Stiger et al., 1998), which applies particularly to GEE. Therefore small sample studies are of special interest. Stiger et al. (1998) analyzed ordinal data in a 2*4 split-plot design with small samples sizes (20, 40, 60, 80) and examined the performance with respect to error rates and power of ANOVA (with and without the Huynh-Feldt-adjustment), MANOVA and GEE. Although a 4-point scale had been used, the results may be adapted to binary data. The ANOVA with adjustment as well as the MANOVA perform well for all sample sizes, while the unadjusted ANOVA behaves sometimes slightly liberal if sphericity was not given. In contrast the GEE exceeds the type I error rates usually for $N<60$. Concerning the power, ANOVA is overall superior, while MANOVA has the lowest rates. Mancl & DeRouen (2001) summarized a number of studies examining the behaviour of GEE in small samples, and concluded that for $N<50$ the type I error rate is generally much too high. McNeish & Stapleton (2016) compared, among other methods, GEE and GLMM for very small sample sizes ($4 \le n_i \le 14$), but unfortunately for a continuous outcome. They found that GEE is generally a poor choice, while GLMM provides satisfying results. McNeish & Harring (2017) confirmed these results for binary variables. In contrast the results of Ma et al. (2012), who also compared GEE with ANOVA, considering continuous and binary variables, and found that GEE keeps the type I error rate even for small N and has the largest power. One shortcoming of all the mentioned studies: in matters of the type I error rates, only the null hypotheses are checked, which hide possible impacts of other factors in the design.

Finally a couple of warnings in this context: „*When applied to modeling binary responses, different software packages and even different procedures within a package may give quite different results*“ (Zhang et al., 2011). „*This kind of convergence problem is a common occurrence in mixed-effects modeling*“ (Fox & Weisberg, 2015). They also report that SAS (Proc NLMIXED) and R (`lme4` and `glmmML`) yield different results for the same datasets, though they all use the same integral approximation approach. By the way, sincere convergence problems are reported in quite a number of publications, e.g. by Beckman & Stroup (2003), who also tell: „*The SAS-available GLMM algorithms considered in this paper performed poorly with fewer than 20 subjects per treatment. ... This raises significant questions about the viability of studies with few subjects and binary data*“.

Unfortunately, as already mentioned before, most studies deal only with one-way designs, except those concerned with GEE and GLMM. Thus there is little knowledge neither of the behavior of the interaction nor of main effects, if there are other nonnull effects. Also, there seems to be no general tendency in favor of one method for either design. So this study tries to fill these gaps. However those methods are focused, which are easily available in the statistical packages and give quick results for the global hypotheses.

3. Methods to be compared

The models, procedures and tests will be presented in the original form, usually for continuous response variables, while they will be applied to dichotomous responses. It will be remarked, for which type of design, between subject or split-plot, they are applicable. More information, especially how to use them in R or SPSS, can be found in Luepsen (2015).

The parametric F-test

In the case of a between subject design 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), normally distributed error e_{ijk} (k=1,..,n_{ij}) with equal variances, 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 $\mathbf{y}^\top = \mathbf{X}\,\mathbf{p}^\top + \mathbf{e}^\top$ using the least squares method, where **y** are the values of the dependent variable, **p** 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). The hypotheses of no effects, e.g. for factor A α_i=0, correspond to equal probabilities p_i in the case of a binary response.

In the case of a mixed design the classical approach will be used (see e.g. Winer et al., 1991), though in recent publications often mixed models, considering e.g. covariance structures, are preferred. For one grouping factor A and one repeated measures factor B, often called trial factor, the 2-factorial ANOVA model for a dependent variable y with $N=\sum n_i$ observations shall be denoted by

$$y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \tau_{ik} + (\beta\tau)_{ijk} + e_{ijk}$$

with α_i, β_j, e_{ijk} as above, n_i subjects per group and a subject specific variation τ_{ik} (k=1,..,n_i). Additionally the covariance matrices are assumed to be spherical and equal for i=1,..,I. The sums of squares and mean squares of the effects are the same as above, if N is substituted by NJ, due to the different definition of N, whereas those for the error terms are different:

$$MS_{between} = \left(J\sum_i\sum_k(\bar{y}_{i\cdot k}-\bar{y}_{i\cdot\cdot})^2\right)/(N-1)$$

$$MS_{within} = \left(\sum_i\sum_k\sum_j(y_{ijk}-\bar{y}_{\cdot\cdot k})^2\right)/(N(J-1))$$

$$MS_{error(between)} = \left(J\sum_i\sum_k(\bar{y}_{i\cdot k}-\bar{y}_{i\cdot\cdot})^2\right)/(N-I)$$

$$MS_{error(within))} = \left(\sum_i\sum_k\sum_j(y_{ijk}-\bar{y}_{i\cdot k}-\bar{y}_{ij\cdot}+\bar{y}_{i\cdot\cdot})^2\right)/((N-I)(J-1))$$

and the F-ratios as

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

To make up for heterogeneous variances, i.e. here unequal p_i on factor B, an appropriate adjustment of the degrees of freedom for the F-test is applied. Here the Huynh-Feldt adjustment, abbreviated H-F, is chosen (see e.g. Winer et al., 1991).

Puri & Sen tests (L statistic)

The tests by Puri & Sen (1985), often referred as L statistic, offer a nonparametric test statistic for the General Linear Model (see e.g. Harwell & Serlin, 1989 and Thomas et al., 1999). In the case of ANOVA models, the hypothesis tested is the identity of distributions. The resulting test statistics are asymptotically χ^2 distributed. They can be seen as a generalization of the well-known Kruskal-Wallis H test (for independent samples). It is well-known that the H test can be performed by ranking *y*, conducting a parametric ANOVA and finally computing χ^2 ratios using the sum of squares (see e.g. Winer,1991). In fact the same applies to the generalized tests. The χ^2-ratios are computed in the case of only grouping factors as

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

and in the case of a mixed design for the tests of A, B and AB as

$$\chi^2_A = \frac{SS_A}{MS_{between}} \qquad \chi^2_B = \frac{SS_B}{MS_{within}} \qquad \chi^2_{AB} = \frac{SS_{AB}}{MS_{within}}$$

Here SS_A, SS_B, SS_{AB}, or generally SS_{effect} , are the sum of squares as outlined before, but computed for R(*y)*, the ranks of *y*, where midranks are used in case of tied values. $MS_{between}$ and MS_{within} are the mean squares previously defined, and MS_{total} 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 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 mean squares in the denominator of the χ^2-tests of Puri & Sen's L statistic increase 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).

Brunner, Munzel and Puri (ATS)

The authors reflect the relative effect of a random variable X_1 to a second one X_2 , which 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 even dichotomous scale, if e.g. $X_1, X_2 \in \{0, 1\}$ (see Noguchi et al., 2012). Based on the relative effect, they 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). In contrary to the WTS, the ATS accounts for the sample sizes that makes it attractive for small cell counts (see Brunner & Munzel, 2002). Both tests check the hypothesis of equal distribution functions, similar to that of the L statistic. For between subject designs detailed descriptions can be found in Brunner & Munzel (2002, chapter 3), Akritas et al. (1997a) as well as in Luepsen (2017). These tests have been extended to repeated measures designs by Brunner et al. (1999b). Bathke et al. (2009) described the procedures, which involve a lot of matrix algebra.

Koch's ANOVA

Gary Koch (1969 and 1970) proposed a couple of nonparametric procedures for split-plot designs based on a multivariate version of the Kruskal-Wallis test and a nonparametric analogue of the one-way MANOVA based on the trace (see e.g. Chatterjee & Sen, 1966). The hypothesis tested: equal mean ranks of the groups considered. This corresponds to equal probabilities p_i in

the case of a binary response. The resulting test statistics are approximately χ^2 distributed. There are several variants for the cases with and without compound symmetry, as well as with and without independence of the factors A and B. The version used here assumes an interaction, but no compound symmetry. A detailed description of the method and the extensive computational procedure can be found in Koch (1969) and shall not be reproduced here.

χ^2 test and log-linear model

For between subject designs Pearson's χ^2 test is performed. The test of the main effects is received from the classical test of independence for two variables. And to test the interaction of factors A and B, a log-linear model including all 2-way interactions is fitted, which yields the desired result. This method requires a sufficient *N*, because for smaller samples too many of the expected cell frequencies may be 1.0 or less. For details see e.g. Agresti (2002). At this point it should be remarked that the χ^2 and the F-test are algebraically similar, and under the null hypothesis asymptotically equivalent, as D'Agostino (1972) showed.

Logistic Regression and Probit Regression

In contrary to the methods above these two are designed for a binary dependent variable with independent observations. Instead of building a model for *y* they model the probability of *y*=1:

logistic regression $$P(y = 1) = \exp\left(\sum_i^P \beta_i x_i\right) / \left(1 + \exp\left(\sum_i^P \beta_i x_i\right)\right)$$

probit regression $$P(y = 1) = \Phi\left(\exp\left(\sum_i^P \beta_i x_i\right)\right)$$

Here x_i (i=1,..,P) are predictors, which correspond in an ANOVA environment to design variables, β_i are the regression parameters, and Φ the normal distribution. The computational procedures are described e.g. in Agresti (2002) and not repeated here. As ML estimation is used, a large *N* is essential. Often 10 per each β_i and a minimum of 100 is postulated (see. e.g. Peng et al., 2002). Primarily β_i =0 is tested by means of a Wald test, or approximately by a t-test. But an ANOVA-type test, into which all β_i belonging to the same effect are summarized, is desirable. It is available by means of a Wald test (see below) or LR (likelihood ratio) test. Usually the latter is preferred (see e.g. Agresti, 2002, and Fox, 1997), especially for smaller samples as analyzed in this study.

Hotelling-Lawley's multivariate test

This test is often used for the analysis of repeated measures designs, because it does not require a compound symmetry of the variance-covariance matrix of *y*. Instead a multivariate normal distribution is demanded. Therefore this test does not seem appropriate for the analysis of dichotomous dependent variables. Nevertheless various authors tried it with differing success (see chapter 2). First the differences of two consecutive measurements are computed $d_{1ik} = y_{i2k} - y_{i1k}$, $d_{2ik} = y_{i3k} - y_{i2k}$, ,.. (for i=1,..,I and k=1,..,n_i). Then d_1 , d_2 ,.. are checked for 0 by means of Hotelling Lawley's test, resulting in an approximately F distributed test statistic (see e.g. Winer et al., 1991), which corresponds to equal differences p_2-p_1, p_3-p_2 ,... in the case of a binary outcome.

GEE (Generalized Estimating Equations)

The GEE method (Liang & Zeger, 1986) can be considered as an extension of the logistic regression to designs with repeated measurements. The specification of the model requires the type of correlation matrix of *y*. Possible correlation structures are among others the compound

symmetry (CS), also often named exchangeable, and the autoregressive (AR(1)). A short sketch of the model for a dichotomous *y*: let

$$P(y_{jk} = 1) = p_{jk} = \frac{\exp\left(\sum_i^P \beta_i x_{ijk}\right)}{1 + \exp\left(\sum_i^P \beta_i x_{ijk}\right)}$$

with k=1,..,N and j=1,..,J, as well as i=1,...,P predictors and corresponding regression parameters β_i. Here, x_{ijk} is the design matrix of subject k, and $V_k = A_k^{1/2} R_k(\alpha) A_k^{1/2}$ with a correlation matrix $R_k(\alpha)$ for $y_k = (y_{k1}, y_{k2},...)$, which can be parametrized by a vector α, and $A_k = \text{diag}(p_{k1}(1-p_{k1}), p_{k2}(1-p_{k2}),...)$. Then the GEE estimates of β_k are the solution of

$$\sum_k D_k V_k (y_k - p_k) = 0 \quad \text{where } D_k = \frac{\partial p_k}{\partial \beta}$$

and $p_k = (p_{k1}, p_{k2},...)$, $\beta = (\beta_1, \beta_2,...)$ (see Emrich & Piedmonte, 1992). McNeish & Stapleton (2016) give a detailed description of the general model and the estimation process. Also to mention Ziegler et al. (1998), who summarize a number of variants and different estimation methods for GEE. The GEE approach is based on LS estimation and produces virtually unbiased estimates, even if the correlation structure is misspecified (see Emrich & Piedmonte, 1992 and Pan & Connett, 2002). On the other side, as the method is based on large sample asymptotic theory, it is not surprising, that for small samples N<50 the tests of the parameters β_k are generally liberal (see e.g. Qu et al., 1994 and Stiger et al., 1998). Responsible is the variance-covariance matrix of β_k, normally computed by means of the sandwich estimator by Liang & Zeger (1986). A number of authors proposed bias-corrected sandwich estimators, among others Fay & Graubard (2001), Kauermann & Carroll (2001), Mancl & DeRouen (2001), Morel et al. (2003), Pan & Wall (2001), Gosho et al. (2014) and Wang & Long (2011). Their work is summarized and compared by Fan et al. (2013), Fan & Zhang (2014) and Wang et al. (2016). However, McNeish & Stapleton (2016) found that GEE is a poor choice for small samples, even combined with one of the above mentioned corrected estimators, except the version by Morel et al. (2003). The hypotheses tested are the same as for the logistic regression.

GLMM (Generalized LinearMixed Models)

Also the GLMM method can be considered as an extension of the logistic regression to designs with repeated measurements. A sketch of the model for a dichotomous *y*:

$$P(y_{ik} = 1) = \frac{\exp\left(\sum_i^P \beta_i x_{ijk} + \sum_i^Q \gamma_{ik} z_{ijk}\right)}{1 + \exp\left(\sum_i^P \beta_i x_{ijk} + \sum_i^Q \gamma_{ik} z_{ijk}\right)}$$

But here, in addition to the fixed effects β_i (i=1,...,P) with design matrix x_{ijk}, there are also random effects γ_{ik} (i=1,...,Q) for subject k with a design matrix z_{ijk}, e.g. for modelling subject and repeated measures effects, and to reflect the correlation among the observations of the same subject, often called cluster in this context. γ_{ik} are multivariate normal distributed with $E(\gamma_{ik})$=0. Similar to the logistic regression an explicit error term e_{jk} is missing (sse e.g. McNeish & Harring, 2017). A correlation structure, as for GEE, has not to be stated here. One advantage of this approach is the flexibility in handling missing data, though such datasets are not considered here. In contrary to GEE, GLMM uses ML estimation methods, which lead to a number of different solutions and programs, e.g. restricted maximum likelihood estimation (REML), Penalized quasilikelihood, Laplace approximation, Gauss-Hermite quadrature or Markov chain Monte Carlo. Details, especially concerning the ML estimation, can be found at

Tuerlinckx et al. (2006) and Song & Lee (2006). Similar to GEE, here also the method is based on large sample asymptotic theory, with the consequence that for small $N<50$ the tests for β_k and γ_{ik} are sometimes liberal. Li & Redden (2015) discuss a number of solutions for this problem, which lies in the estimation of the denominator degrees of freedom (*ddf*) for the F-test, into which the Wald test is transformed. The most popular solution is probably the rather complicated one by Kenward & Roger. The most simple one uses $ddf=N$-rank(C), where C is the contrast matrix. Additional ANOVA-like tests are mentioned below. The hypotheses tested are the same as for the logistic regression.

Wald tests

The primary results from an analysis using logistic regression, probit regression, the GEE or GLMM method are the estimates of the model parameters β_i together with their standard errors and a significance test of $\beta_i=0$ for each i, normally by means of a Wald test. But in this context an ANOVA-like test is desired, into which all β_i belonging to the same effect are summarized. On one side there is Wald's χ^2 test in the variant for several parameters (see e.g. Carr & Chi, 1992 and Pan & Wall, 2001):

$$(C\hat{\beta})'(CV_{\beta}C')^{-1}(C\hat{\beta})$$

which is approximately χ^2 distributed with *rank(C)* degrees of freedom, and where $\hat{\beta}$ are the estimates of β, V_β is the variance-covariance-matrix of β, and C a contrast matrix, and in its simpler form (see e.g. Kenward & Jones, 1992)

$$\hat{\beta}'V_{\beta}^{-1}\hat{\beta}$$

where $\hat{\beta}$ and V_β are restricted to those i belonging to the effect of interest. Fan & Zhang (2014) found that the above test is too liberal for small sample sizes and proposed a different one, based on the work of Akritas et al. (1997a) and Brunner et al. (1997):

$$Q = \hat{\beta}'C(C'C)^{-1}C'\hat{\beta}$$

The expression $(c_1/c_2)Q$ is approximately χ^2 distributed with f degrees of freedom where

$$c_1 = tr(TV_{\beta}) \qquad c_2 = tr(TV_{\beta}TV_{\beta})$$

$$f = c_1^2/c_2 \qquad T = C(C'C)^{-1}C'$$

Fan & Zhang (2014) showed in their study of GEE for repeated measures models with $5 \le n_i \le 20$, that their ANOVA-type test is able to control the error rates in most situations, while the Wald test produces rates up to 80% for the trial effects.

While the Wald test above is equivalent to a Type III test, Fox & Weisberg (2015, chapter 4.4.4) favored a Type II Wald test which is offered in the function `Anova` of the R package `car`. It is based on the likelihood ratio method, using analysis of deviance tests. This one conforms to the principle of marginality and is most powerful in the case of no interaction. Using it, the main effects may be overestimated, in contrary to the interaction effects.

4. The Study

The aim of this study is to identify one or a couple of methods, which allow the analysis of a binary response in a factorial ANOVA layout. For this reason the impact of several settings of such a design on the type I error rates and the power is investigated by means of a Monte Carlo study with 2000 replications. These settings are the type (between subject, split-plot), size

(number of cells), cell frequencies (equal, unequal), cell counts (5,10,...,50), pairing (positive, negative), effect of factors and interaction, binomial probabilities (p=0.1, 0.2, 0.5, 0.8, 0.9) and correlation structure (equal or unequal correlations). This should cover all important situations and allow for generalizations. The resulting sample sizes N vary from 10 to 1000. Without loss of generality the layout will be restricted to two factors A and B, and for each factor only one vector of effect sizes has been chosen, which should suffice to see, if one factor has at all an impact on the results. In the case of mixed designs A shall denote the grouping factor, B the trial factor and AB the interaction. p denotes the overall fraction of the binary outcome and p_i the corresponding values for the groups of A.

There are two major designs: a between subject and a mixed (split-plot) design. For both the following subdesigns are analyzed:

- a 2*4 design ("small design") with equal cell counts (balanced) and one with unequal cell counts and a ratio $\max(n_{ij})/\min(n_{ij})$ of 3 (unbalanced), and
- a 4*5 design ("large design") with equal cell counts (balanced) and one with unequal cell counts and a ratio $\max(n_{ij})/\min(n_{ij})$ of 4 (unbalanced).

The binomial probabilities p have been set to 0.5, 0.8 and 0.9 (equivalent to 0.5, 0.2 and 0.1), as for $0.25 \leq p \leq 0.75$ the variances of a binomial distributed outcome can be regarded as equal. For the split-plot design the following correlation structures have been chosen which are assumed equal for all groups:

- exchangeable (equal covariances, compound symmetry) with r=0.3, a value that seems realistic and had often been chosen (see e.g. Emrich & Piedmonte, 1992), and
- descending correlations r=(0.7, 0.5, 0.4, 0.2) which is similar to the AR(1) structure and denoted as ar1 (unequal covariances, no sphericity or compound symmetry).

In the case of between subject designs, noting that A and B are exchangeable, the type I error rates of the main and interaction effects had been checked for the case of a null model, the case of one significant main effect (A(0.6) or B(0.6)), and the case of a significant interaction AB(0.4). In the case of mixed designs the type I error rates of all main and interaction effects had been checked for the case of the null model, the case of one significant main effect A(0.6) or B(0.4), and the case of a significant interaction AB(0.4). Here e.g. A(d) denotes an effect of size d for factor A, corresponding to effect vectors $\boldsymbol{p}^{\mathsf{T}} + \left(-s\frac{d}{2}, 0, \ldots, 0, s\frac{d}{2}\right)^{\mathsf{T}}$, where $\boldsymbol{p}$ =(p,..,p) with the overall probability p and $s=\sqrt{p(1-p)}$ its standard deviation. Analog definitions for B(d) and AB(d). In some instances additional design sizes and correlation structures were analyzed for selected models, in order to assure some of the results.

For unbalanced designs the interaction effects $(ab)_{ij}$ had to be adjusted respecting the different cell counts, in order to avoid impacts on the main effects. It should be remarked that most ANOVA procedures are based upon LS estimation, which corresponds to weighted means analysis, where the cell counts n_{ij} have a larger impact on the results than with the unweighted means analysis. The latter assumes equal cell counts by design, and allows only a couple of missing observations (see also Winer, 1991). Unfortunately the ATS method for split-plot designs, as implemented in the R package `nparLD`, is based on the unweighted means analysis (see Noguchi et al., 2012), which may lead to results, which are not comparable with those from the other analyses.

Unfortunately first simulations revealed a failure of the data generation in mixed design models,

when for p=0.9 in one factor level the effect had to be added: p_i =0.9+s*d/2 (see above). In order not to let the shifted parameter p_i come too close to 0 or 1 respectively, the p had to be reduced generally to p=0.88. The problems intensified in the case of an unequal correlation structure with descending correlations (ar1), where additionally effect sizes had to be reduced from 0.6 to 0.4 (for factor B) and 0.4 to 0.36 (for the interaction). Even worse was the case with two non-null effects, e.g. for the analysis of the power when there are also other effects present. Then the effect sizes had to be scaled down to 0.3.

Another problem to be investigated is the pairing of n_i and p_i. Being aware that for p>0.5 the variances of y become smaller with increasing p, it is to be expected, that the F-test reacts liberal, if levels of A with larger n_i have also larger p_i, and that it reacts conservative, if levels with larger n_i have the smaller p_i. Of course, the same behavior will apply to the case p<0.5. Therefore the effects of factor A will be analyzed for all three relations of n_i and s_i: independent, positive and negative pairing. Finally, as in the case of p=0.5 with an effect d for factor A, the resulting p_i-s*d/2 and p_i+s*d/2 will be equal and therefore produce equal variances, p=0.6 is chosen instead, when situations of heterogeneity are analyzed.

The type I error rates (at 5% and 1%) and the power were computed for n_i=5,10,15,..,50 as percentages of rejected null hypotheses. Although generally 2000 replications were chosen, for the GEE and the GLMM methods the number of repetitions have been limited to 1000 because of the enormous computational effort. The relatively small number of samples is not unusual (see e.g. McNeish, 2017 and Guerin & Stroup, 2000). Due to the convergence problems, mentioned in the previous chapter, which occurred mainly with GLMM in smaller samples $n_i \leq 15$, the actual number of repetitions has been reduced sometimes by about 2 percent. But the situation became much worse with GEE, which produced unmanageable covariance matrices for smaller samples $n_i \leq 10$. The failure rates reached sometimes 90 percent. In those cases the repetitions had to be increased to 5000, in order to receive at least 200 valid results, or the sample size of 5 had to be dropped from the study, especially for unbalanced designs.

De facto the study ran in two parts: in a first step all methods mentioned in chapter 3 were examined, but only for two designs: small balanced and large unbalanced. Depending on the results and on the evaluation by other authors (see chapter 2), some methods have been dropped from the main study in the second step. For the between subject design these were the log-linear model and the probit regression. The log-linear model, because the type I error rate increased beyond 0.10 (for α=0.05) in many situations, which had to be expected from Swafford's study (1989), and because most studies prefer the F-test instead, and the probit regression, because most authors see advantages for the logistic regression. For the logistic regression, a compromise test has been chosen as ANOVA-like test, composed by the χ^2-values of the LR and the Wald test with the same degrees of freedom, denoted by WLR:

$$\chi^2_{WLR} = (\chi^2_{Wald} + \chi^2_{LR})/2$$

The reason: especially for small samples, the LR test behaves rather liberal, while the Wald test acts extremely conservative. Concerning split-plot designs, only the GEE method has been dropped (for more details see below).

For the GEE and GLMM analysis it was necessary to select a suitable method and function in the preliminary study to apply them in R. For the GLMM analysis all three estimation methods together with the Wald tests mentioned in chapter 3 were compared. The only satisfying procedure was REML (R function `glmer`), which held the error rates under control on the whole. But, unexpectedly, it is the Type II Wald test, which managed also the case of a significant in-

teraction. In contrast, the other ANOVA-type tests as well as the other two GLMM methods revealed exploding error rates with increasing sample sizes. Therefore GLMM in conjunction with the Type II Wald test is chosen for the main study, supported by the positive judgements by McNeish & Stapleton (2016), McNeish & Harring (2017), Oberfeld & Franke (2012) and Jaeger (2008), and despite the computational problems cited previously and confirmed in the preliminary step.

For the choice of the GEE procedure, the focus has been laid upon the different estimation methods for the covariance matrix $V_{\hat{\beta}}$ of the parameter estimates $\hat{\beta}_i$. First, as a basis for the estimation of the parameters themselves, the method by Prentice & Zhao (1991) was applied. All 9 methods described by Wang et al. (2016) were compared. In general the solutions from Pan & Wall (2001), Gosho et al. (2014) and Wang & Long (2011), which obtain their estimates by pooling observations across different subjects, as well as the method by Morel et al. (2003), have the most benevolent behavior. As to be expected: the ANOVA-like tests by Fan & Zhang (2014) show generally much smaller error rates than the Wald test, but with the disadvantage of an also much smaller power. For the first three methods additionally the ANOVA-type test by Pan (2001) was computed, which is able to control the type I error rate in a same way as the one by Fan & Zhang (2014), but shows on the other side clearly better power rates. The error rates and power for the previously mentioned four methods, together with the one by Liang & Zeger, and applied to the ANOVA-type tests by Wald, Fan & Zhang and Pan are to be found in appendixes B9 and B10. These show that the type I error rates rise sometimes up to over 50 percent (see 7.3 and 7.6 in B9), even for the best performing GEE methods and ANOVA-like tests. As a consequence from these experiences, the computational problems with the estimation of the covariance matrices, and the observation that GEE tends to exceed the type I error rates for small samples (see chapter 2), this method has been dropped from the main study.

Computational aspects concerning the data generation and the selection of ANOVA procedures are to be found in the last chapter.

5. Results

Tables and Graphical Illustrations

The following remarks represent only a small extract from the numerous tables and graphics produced in this study and will concentrate on essential and perhaps unexpected results. All tables and corresponding graphical illustrations are available online (see address below). These report the proportions of rejections of the corresponding null hypothesis, for different models and n_{ij} = 5,10,..,50. They are structured as follows:

Results from the main study (α=0.05)
(for all methods considered, in 2*4 and 4*5 as well as balanced and unbalanced designs):

- B 1: type I error rates for fixed n_{ij} in between subject designs,
- B 2: power in relation to n_{ij} in between subject designs,
- B 3: type I error rates for fixed n_i in mixed designs,
- B 4: power in relation to n_i in mixed designs,

Results from the preliminary study (generally at α=0.05, some also at α=0.01)
(for all methods in 2*4 balanced and 4*5 unbalanced designs):

- B 5: type I error rates of all methods for fixed n_{ij} in between subject designs,
- B 6: power of all methods in relation to n_{ij} in between subject designs,
- B 7: type I error rates of all methods for fixed n_i in mixed designs,

- B 8: power of all methods in relation to n_i in mixed designs,
- B 9: type I error rates of selected GEE methods for fixed n_i in mixed designs,
- B 10: power of selected GEE methods in relation to n_i in mixed designs,

All references to these tables and graphics will be referred as B *n.n.n*. All tables and graphics can be viewed online: http://www.uni-koeln.de/~luepsen/statistik/texte/comparison-tables/. A note to the figures which show the behavior of the type I error rates: first they have been smoothed using moving averages over the range of n_i=5,..,50 to suppress spurious values, then the maximum of the 10 values has been chosen.

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 - can 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. In this study Peterson‘s moderate robustness will be applied, i.e. an acceptance interval [3.75 , 6.25]. As the results concern the error rates for 10 sample sizes n_{ij} = 5,...,50, it seems reasonable to allow a couple of exceedances within this range. The following remarks concern the results for tests at α=0.05. As noted in several other studies (see e.g. Luepsen, 2017) nearly all tests behave more liberal at α=0.01. Concerning the binomial probabilities p the values 0.8 and 0.9 are used, reminding that these are equivalent to 0.2 and 0.1.

5. 1 Results I: between subject designs

The most exciting question is: how behaves the parametric F-test in those cases which were not treated by Lunney (1979)? These are small samples ($N \leq 20$), unbalanced designs and the influence of nonnull effects of other factors. The control of the type I error rate is guaranteed, even for small N=10, as long as $0.3 \leq p \leq 0.7$, while for extreme $p \geq 0.8$, respectively $p \leq 0.2$, the rates for the interaction effect rise up to 7 (for p=0.8) and 10 (for p=0.9), if there is a nonnull main effect (see table 2 and appendix table B 1.7 and 1.9). And this occurs even for balanced designs. This accords with the requirement of the classical ANOVA for safe tests: equal variances. A more detailed inspection revealed, that this is mainly due to the larger number of cells, as it occurs only for large designs (see figure 2). More severe violations occur in unbalanced designs, if n_i and p_i are dependent and $p \geq 0.8$, the case of negative or positive pairing. Here the type I error rates for the test of a main or interaction effect do not lie any longer in the interval of robustness. E.g. in the case of a significant factor A, even for a small ratio $\max(n_i)/\min(n_i)$ of 1.3, the rates for the test of main effect B rise to nearly 12 (see table B1.3), and to 15 for the test of interaction AB, if n_i and p_i are positively correlated, and fall to 2, if they are negatively correlated (see B 1.8). Similar results are obtained for the tests of the main effects, if the interaction AB is significant (see B1.5 and figure 1). It should be remarked that the violations are independent of the cell frequencies 5,...,50.

Considerably better performs Puri & Sen‘s L statistic in these situations of heterogeneity, because it exceeds only for p=0.9: with values near 7 (see B 1.3, 1.8 and 1.5) for the tests of the main effects and with values near 9 (see B 1.8 and 1.10) for the tests of the interaction. The only method that keeps the type I error rate without exceptions, is Brunner & Munzel‘s ATS. Both results confirm the findings cited in chapter 2. The logistic model together with the proposed WLR-test is able to control the type I error, except in one situation: if there is a nonnull interaction effect. Then the error rates rise up to values between 10 and 20, mainly in unbalanced de-

signs (see B 1.4 and 1.5). Additionally there are serious exceedances of the error rate for the test of the interaction, if both main effects are significant. These models have not been considered in the literature, and therefore the rating of the LR diverges here. The results suggest that the logistic approach is no alternative to the F-test.

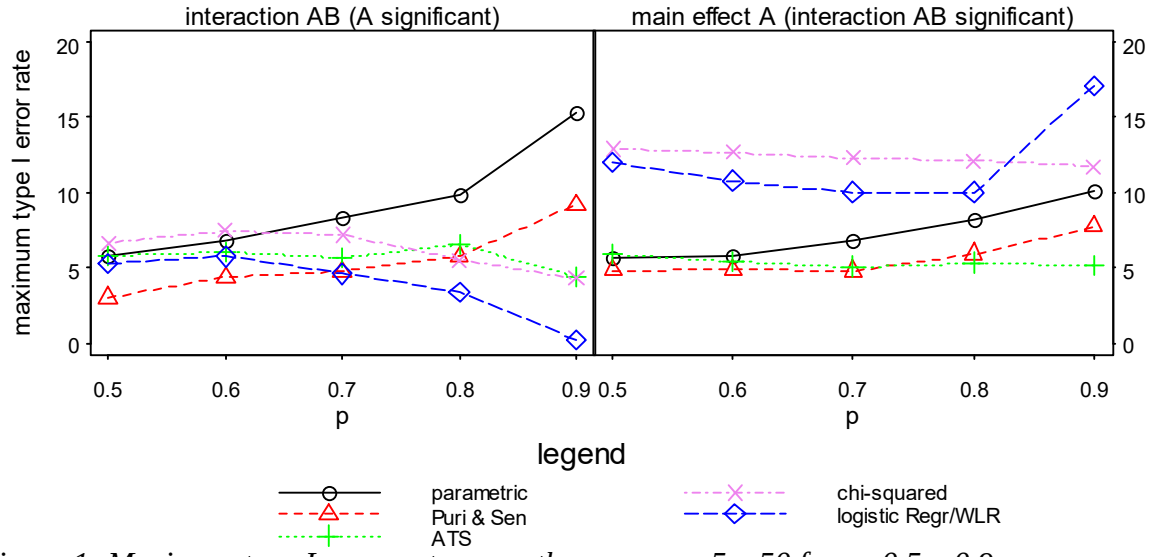

Figure 1: Maximum type I error rates over the range n_i=5,..,50 for p=0.5,..,0.9: for the test of the interaction AB when one main effect is nonnull (left), and for the test of a main effect if the interaction AB is nonnull (right), both in between subject designs, when n_i and p_i are not independent.

Concerning the power, the parametric F-test is always among the best performers. Puri & Sen's L statistic has nearly identical rates. Only for the test of the interaction the rates lie sometimes below those of the F-test for small n_i: about 20% for n_i=5 and about 10% for $n_i \leq 15$ (see B 2.5). Apparently the L statistic performs better for binary variables than for metric outcomes when comparing these results with those cited by Sawilowski (1990). The ATS is able to keep up only in balanced and small unbalanced designs, whereas in large unbalanced designs the rates are by far the lowest, mostly for smaller samples: for n_i=5 (p=0.5), for $n_i \leq 20$ (p=0.8) and $n_i \leq 30$ (p=0.9). Here the rates lie between 20% and 60% below those of the F-test (see e.g. B 2.1 and 2.2). This reassures the results cited in chapter 2, among others by Richter & Payton (1999). Compared with the other methods, the logistic model exhibits in all situations an unsatisfying performance in respect to the power.

5. 2 Results II: mixed designs

Split-plot designs, as mixed designs are often called, require a more detailed analysis, because first, the factors A and B are not exchangeable, and secondly, the correlation structure of the repeated measurements has to be taken into account. At first the case of equal correlations will be regarded.

The F-test controls the type I error fairly well, in balanced and unbalanced designs, but with the exception of the interaction effect, if there are also other nonnull effects. Together with a non-null trial effect, the test of AB reacts slightly liberal for extreme p=0.9 with rates between 7 and 8 (see B 3.11.). In this case the Huynh-Feldt adjustment is to prefer. A similar behavior can be observed, if there is a grouping effect with independent p_i and n_i. But here the multivariate test is the only alternative (see B 3.9.1). A detailed look into the results exhibits again, that mainly the size of the design is responsible for the violations cited above. Things look worse, if in

unbalanced designs the grouping factor A has a nonnull effect, which leads to heterogeneous variances, and the p_i coincide with the n_i, the case of negative pairing. For positively correlated n_i and p_i the type I error of the F-test shows rates between 15 and 20, rising with increasing p 0.5 -> 0.9 (see figure 3), and for negatively correlated n_i and p_i rates below 2 (see B3.10). And unfortunately neither the Huynh-Feldt adjustment, nor Hotelling-Lawley's multivariate test, nor Koch's procedure are able to reduce the rates clearly for $p > 0.6$. Here also the violations are independent of the cell counts 5,...,50. In these situations the only methods, which remain completely unaffected, are the ATS and GLMM (see figure 3).

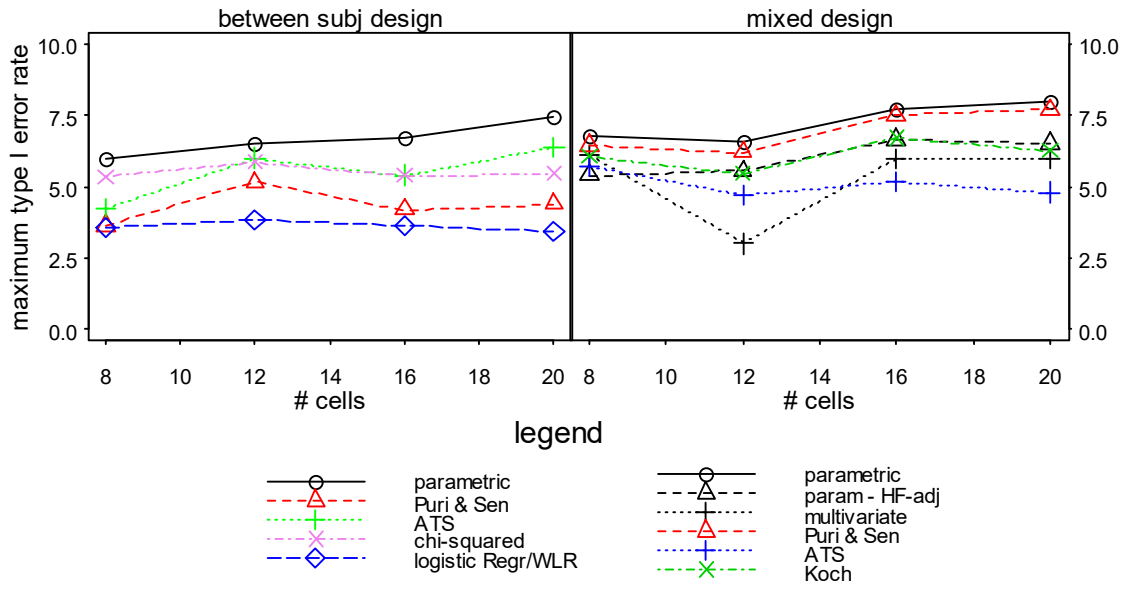

Figure 2: Maximum type I error rates over the range n_i=5,..,50 vs. the number of cells (8,.., 20) for the test of the interaction AB in a balanced design, if there is of a nonnull grouping factor A, for p=0.8, in between subject designs (left) and mixed designs (right).

A look onto the other methods: Puri & Sen's L statistic behaves very similar to the parametric F-test, with the advantage of somewhat lower error rates, though still beyond the limit of moderate robustness in the cases mentioned above. While several authors observed a sensitive reaction to unequal variances (e.g. Feir & Toothaker, 1974), the findings here show, that the violations of the type I error are essentially independent of p, and therefore independent of the heterogeneity. Whereas for the above noted methods predominantly only the tests of the repeated measurements effects react sometimes too liberal, it is vice versa with the ATS: only for the test of the grouping effect A the error rates pass beyond the limit of robustness, with values up to 12, but mainly for $n_i \leq 30$ (see B3.1 to 3.3). Here the violations are more severe for unbalanced than for balanced designs. The authors themselves (Brunner et al., 1999a) reported this problem. But unfortunately Brunner & Munzel's test has a problem with interaction effects, as it is based on the unweighted means analysis as mentioned before. This leads to a dilemma when analyzing unbalanced, mainly small designs: if the ATS shows significant results for the trial main and the interaction effect, it cannot be excluded that the outcome for B is due to a nonnull interaction effect. In contrast, the test of the grouping factor is not affected by the interaction. In addition the results of the ATS for the affected tests of the trial factor are issued based on unadjusted interaction effects ab_{ij} (labelled in the tables as ATS uncorr), which exhibit the complete control of the type I error.

Also the GLMM model has deficiencies relating to the type I error: if one factor has a nonnull effect, the main effect of the other factor shows sometimes increasing rates (7-10) for n_i -> 50

for unbalanced designs (see tables B 3.2 and 3.5). Additionally there are some violations, with rates near 7, for the tests of the interaction, but surprisingly only for p=0.5 (see tables B 3.8 and 3.9). On the other side: the GLMM never exceeds the type I error for correlated n_i and p_i, though the rates stay below 2.5 for $n_i \leq 20$. The experiences made here cover overall the results by McNeish & Harring (2017) and other authors cited in chapter 2. As can be concluded implicitly from the remarks in the previous paragraph, the H-F, the multivariate test as well as Koch's procedure show a perfect type I error performance, except for the interaction effect if n_i and p_i are positively correlated. Apart from the last remark, the findings here, a sometimes liberal F-test and correct tests by Huynh-Feldt, Hotelling-Lawley and Koch, cover most of the results from the literature cited in chapter 2, e.g. those by Oberfeld & Franke (2013) as well as Stiger et al. (1998). However the outcome here presents a better type I error control of the multivariate test, especially for the case of unequal correlations. Finally one phenomenon concerning all tests, except GLMM, in small balanced designs: for very small samples (n_i=5) and extreme p (0.9) the tests react extremely conservative, mostly with rates below 2.

At this point it should be reminded of the disappointing type I error control of the GEE method. Even the best performing procedures by Gosho et al. and by Wang & Long in conjunction with the ANOVA tests by Wald or Pan (see chapter 4) are not able to keep the type I error rate in an acceptable range for the test of a main effect, if there is a nonnull interaction, a model that has been rarely included in other simulation studies. In fact, for all methods and ANOVA-like tests the error rates rise up to over 50 percent for ni=50 (see 7.3 and 7.6 in B9), even for equal n_i.

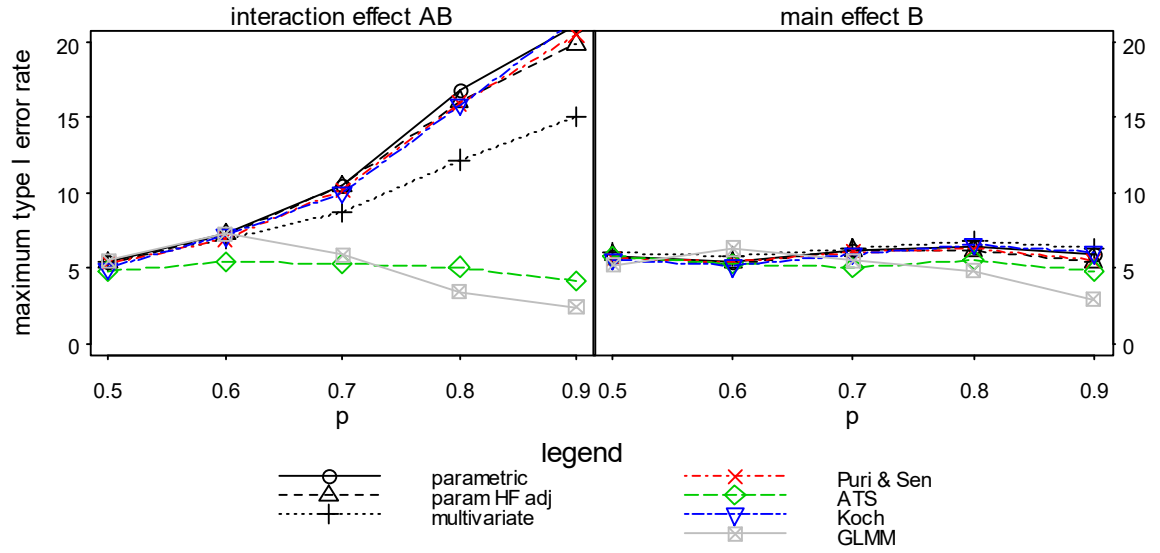

Figure 3: The maximum of the type I error rates over the range n_i=5,..,50 for p=0.5,..,0.9: for the test of the interaction AB (left) and for the test of main effect B (right), both in mixed designs, when the effect of grouping factor A is nonnull and when n_i and p_i are not independent.

Concerning the power, also for mixed designs the parametric F-test is always among the best performers. The power of the other methods will be related to that of the F-test, where only those situations are of interest, where the considered procedure does not offend the type I error seriously. Puri & Sen's L statistic as well as Huynh Feldt's correction for the tests of the trial effects can keep up with the parametric F-test in all models and situations. In contrary, Hotelling-Lawley's multivariate test has often a power superior to the F-test, especially for large designs, e.g. for A and AB with rates of up to 50% higher than those of the F-test (see e.g. B 4.1.2 and B 4.3.2), but sometimes also a power clearly smaller than that of the F-test, e.g. for B with rates of 50% below (see e.g. B 4.2.2), but mainly for small $n_i \leq 20$. Occasionally the power

achieves only 10% of that of the F-test for n_i=5, which covers the findings of Stiger et al. (1998) in part. Generally the multivariate test performs best for $n_i \geq 15$ and in large designs (see e.g. figure 4), which has been remarked also by other authors. Because of the problems listed in the previous paragraph, the ATS is only of interest for correlated n_i and p_i, though its power rates lie often up to 50% below those of the parametric tests (see e.g. B 4.4). Koch's procedure performs very similar to the multivariate test, which is not surprising, because it is based on a nonparametric MANOVA. It can keep up with the F-test in many situations, especially for the test of A, while for B and AB it needs larger samples ($n_i \geq 15$). Finally, the GLMM has a disappointing power on the whole, though with a couple of exceptions: e.g. for $0.3 \leq p \leq 0.7$, especially for the interaction effect. One remark concerning all methods: in small designs the power for A and AB is about 20-30% higher in the case of equal n_i than in the case of unequal n_i, whereas in large designs the rates are rather similar.

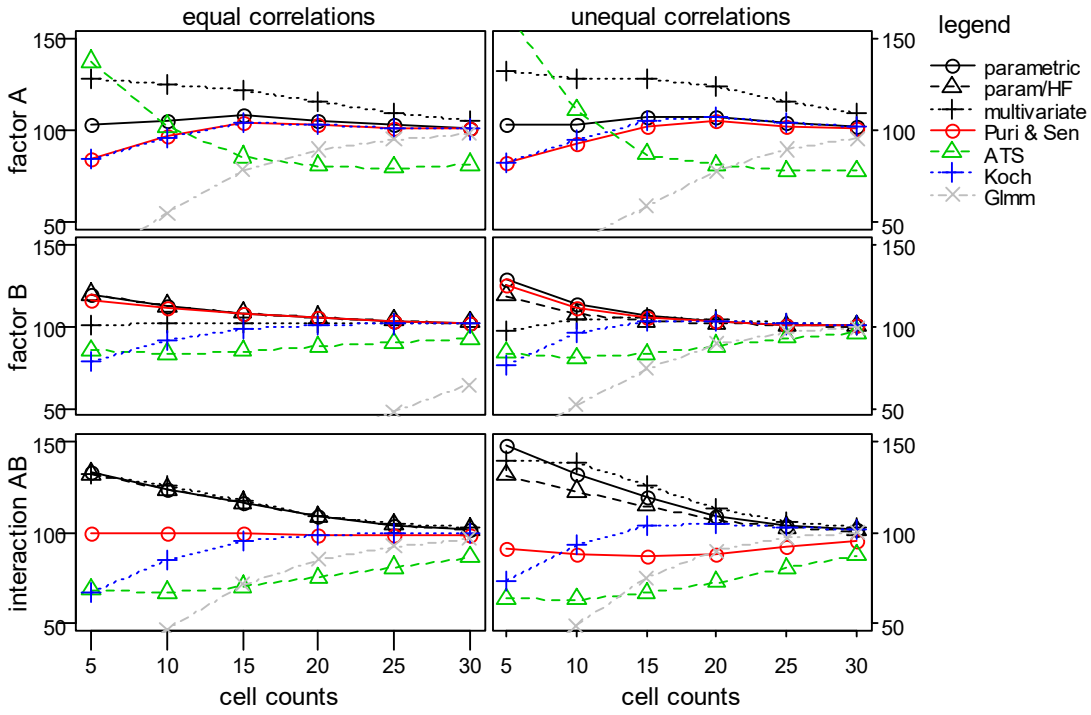

Figure 4: Relative power computed as the percentage of the mean power averaged over the 7 methods in the range of n_i=5,..,30 for factors A and B and the interaction AB, for p=0.5 and large unbalanced designs, with equal and unequal correlations of the repeated measurements, showing the good performance of H-F and the multivariate test, the lower rates of Koch's procedure for small n_i and the poor overall performance of the ATS.

An interesting question might be: how large is the effect of unequal correlations of the repeated measurements in split-plot designs? The parametric F-test and the L statistic show about 15-20% higher type I error rates, and therefore also more violations (see figure 5). This had been observed previously by Harwell & Serlin (1994). Also in this case the interaction effects are affected, even for $p \geq 0.8$, with rates up to 11. It occurs only in large designs, as in the case of equal correlations. In most circumstances Huynh Feldt's correction, the multivariate or Koch's procedure are preferable. However, in the case of a nonnull grouping effect, the multivariate test is the only alternative. Also in the instance of positively correlated n_i and p_i, the results are identical to those for the case of equal correlations. Generally, the ATS, Koch's test and GLMM exhibit no tendency concerning the error rates. Harwell & Serlin reported also a decreasing power for raising covariance heterogeneity. This study confirms this only partly: unequal correlations lead to a loss of about 20% of power for all methods, but only for the test of A, and for the test

of AB if A has a nonnull effect. In contrast, there is a reverse impact on the power for B. Here the F-test shows about 10-20% higher rates. This corresponds to its behavior concerning the type I error in these situations. For the interaction AB, there is no definite tendency observable, if A has no effect (see also table 1). Finally it should be remarked, that the results are very similar for two other correlation structures, which have been examined for a selection of models: ascending correlations r=(0.2, 0.4, 0.5, 0.7) and unstructured correlations r=(0.2, 0.6, 0.1, 0.4).

6. Conclusion and practical aspects

In between subject designs the F-test has complete control of the type I error only if $0.3 \le p \le 0.7$. Even for p=0.8 and equal cell counts one has to accept slight exceedances. The better selection is Puri & Sen's L statistic, which controls the type I error in nearly all situations. Another argument in favor of the L statistic is the overall excellent power, at least for $n_i \ge 15$, and in most cases even for $n_i \ge 10$. Although Brunner & Munzel's ATS has a complete control of the type I error rate, it is no good choice because of its poor power. And finally, the logistic regression has unacceptable error rates rising up to 10 and beyond (n_i->50) in a couple of situations listed in the previous chapter. This makes this procedure, which was made especially for binary variables, a dangerous choice. All in all Puri & Sen's L statistic seems to be the best overall recommendation.

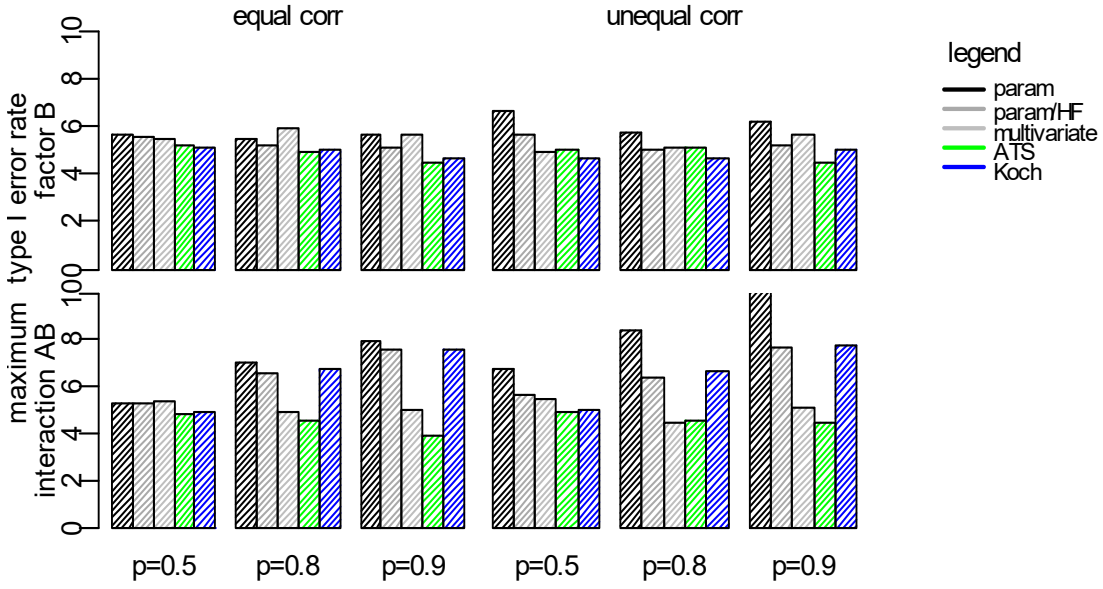

Figure 5: Maximum type I error rates over the range n_i=5,..,50, with equal and unequal correlations of the repeated measurements, in large unbalanced designs, for the effects of factor B and interaction AB, both with nonnull effects of factor A, showing particularly the larger rates for the F-test in case of unequal correlations and the robustness of the multivariate test.

Also in mixed designs the F-test has complete control of the type I error only if $0.3 \le p \le 0.7$, or if the design is balanced with a small the number of cells (≤ 15). On the other hand, it is only the test of the interaction for which the F-test cannot control the error rates. As a consequence, in some situations, e.g. in unbalanced designs with p outside of the interval [0.3 , 0.7], other methods should be chosen. As long as n_i and p_i are not correlated, either the H-F adjustment, the multivariate test or Koch's ANOVA are a good choice, whereas Puri & Sen's L statistic reacts slightly liberal. Nevertheless there is one situation, where the multivariate test is the only

acceptable alternative: for the test of AB in large designs, if $p \geq 0.8$ and A has a nonnull effect. Regarding the power, the H-F can keep up with the F-test on the whole the best, particularly for small $n_i \leq 10$. Now to the challenge of positively correlated n_i and p_i. The case $0.3 \leq p \leq 0.7$ can still be handled by the H-F, multivariate and Koch's procedure (see B 3.6, 3.10.1 and 3.10.2). For $p \geq 0.8$ the only methods without problems with the test of AB are ATS and GLMM. As their power differs considerable, the ATS should be preferred. Because of the poor power of the ATS, with a loss of about 50%, its use should be restricted to this condition.

Beside this, the F-test behaves generally more liberal in large designs (with more than about 15 cells) in the case of between subject designs, but even more in mixed designs (see figure 2), especially for the tests of the interaction effect (see e.g. table 2). But in these situations there are recommendable alternatives: the L statistic in grouping designs, and the multivariate ANOVA with its superior power in larger split-plot designs (with the restriction of $n_i \geq 10$) or the Huynh-Feldt adjustment for the F-test, which achieves the power of the F-test (for small n_i).

The final recommendation, first for between subject designs: if the relative frequencies p of the two values of y lie within the interval [0.3 , 0.7], the parametric F-test may be used without risk. For values outside this range Puri & Sen's L statistic should be the choice, even for equal cell frequencies. For mixed designs there is no unique method to recommend. The F-test is an appropriate choice, if either the frequencies of y lie within the interval [0.3 , 0.7], or if a balanced design with a maximum of 15 cells is the basis. In addition, the F-test may always be applied for the test of the main effects. On all other occasions either Huynh-Feldt's adjustment for the F-test, Hotelling-Lawley's multivariate test or Koch's ANOVA is recommended, with a preference for the H-F in small designs and for small samples $n_i \leq 10$ and for the multivariate approach in case of large designs. But with the following exception: if the relative frequencies of y for the levels of A are not equal, larger than 0.8 (respectively smaller than 0.2) and positively correlated with n_i, then either the ATS or the GLMM should be applied for the tests of the interaction AB, with a preference for the ATS.

7. Programming

This study has been programmed in R (version 3.3.2 and later 3.3.3). For the data generation two different functions had been applied: `runif` in the case of between subject designs to generate uniform distributed data, which were split into two groups at the desired cutpoint p, and `rmvbin` from the package `bindata` for split-plot designs (see Leisch et al. 1998), which is based on the generation of multivariate correlated normal samples and allows the creation of binary variables with specified percentages of p_i and specified correlations.

Various functions had been chosen to analyze the simulated data: the function `aov` in combination with `drop1` (to receive type III sum of squares estimates in the case of unequal cell counts) for the standard ANOVA F-test, an own function `np.anova` for the factorial Puri & Sen-tests, also an own function `ats.2` for the ATS method in between subject designs - meanwhile an appropriate package `GFD` is supplied in R, and the function `nparLD` from the package `nparLD` for ATS in mixed designs. The logistic and probit regression had been performed with `glm`, the χ^2-tests with `chisq.test` and `loglin`, and the multivariate Hotelling-Lawley test with the functions `lm` and `anova`. For Koch's nonparametric analysis of a split-plot design again an own function `koch.anova` had been chosen. For the analysis of GLMM models the following functions had been applied: `glmer` (R package `lm4`), which is based on restricted maximum likelihood estimation (REML) using a bounds constrained quasi-Newton method (`nlminb`, by means of R function `optimx` from package `optimx`), `glmmPQL` (R package `MASS`), which uses Penalized quasilikelihood estimation, and `glmmML` (R package `glmmML`), which applies

adaptive Gauss-Hermite quadrature. For the analysis of GEE models the function `geeglm` (R package `geepack`), based on the estimation method by Prentice & Zhao (1991), had been applied for the parameter estimation. Additionally, the functions from the package `geesmv` had been used to estimate the covariance matrix according to the 9 methods described in Wang et al. (2016), however with a modification, in order to handle failures in the estimation process. For the own functions see Luepsen (2014).

Some of the computations had been performed on a Windows notebook, but for the major part the high performance cluster CHEOPS of the Regional Computing Centre (RRZK) of the university of Cologne had been used. I would like to thank the staff of the RRZK for their technical support as well as Prof. Unkelbach for his organizational support.

effect model	parametric		Puri & Sen		ATS		logistic	
	small	large	small	large	small	large	small	large
A	5.41	5.50	5.94	5.27	5.29	5.27	3.94	4.04
B (A sig, n_i and p_i indep)	5.62	5.31	3.66	3.45	4.62	5.10	4.37	6.54
A (AB sig)	5.95	5.42	4.82	3.96	5.62	4.80	25.65	12.31
AB	5.52	5.59	5.67	5.36	5.17	4.84	4.12	3.52
AB (A sig, n_i and p_i indep)	6.00	9.09	3.98	5.60	3.34	5.44	0.10	0.29
AB (A sig, B sig)	7.12	8.45	5.50	5.84	5.85	5.80	4.65	5.09
	$n_i \sim p_i$	$n_i \mid p_i$	$n_i \sim p_i$	$n_i \mid p_i$	$n_i \sim p_i$	$n_i \mid p_i$	$n_i \sim p_i$	$n_i \mid p_i$
B (A sig, n_i and p_i dep)	10.94	3.24	7.07	1.75	4.74	4.67	6.04	4.41
AB (A sig, n_i and p_i dep)	14.90	5.29	9.18	2.84	4.15	4.24	0.22	0.12
	equal	unequal	equal	unequal	equal	unequal	equal	unequal
A	5.29	5.50	5.27	5.94	5.27	5.29	4.04	3.80
B (A sig, n_i and p_i indep)	5.62	5.28	3.66	2.94	5.10	4.62	6.54	5.61
A (AB sig)	5.42	5.95	4.35	4.82	4.89	5.62	12.93	25.65
AB	5.52	5.59	5.45	5.67	5.17	4.33	4.12	2.77
AB (A sig, n_i and p_i indep)	8.94	9.09	5.60	5.23	5.41	5.44	0.29	0.27
AB (A sig, B sig)	7.12	8.45	5.50	5.84	5.85	5.80	4.65	5.09

Table 2: Maximum smoothed type I error rates for all methods in all situations (between subject designs and p=0.9), for small and large designs plus balanced and unbalanced designs with independent n_i and p_i, as well as for negative pairing ($n_i \sim p_i$) and positive pairing ($n_i \mid p_i$) in large designs.

effect model	corr	p	parametric		param/HF		multivariate		Puri & Sen		ATS		Koch		GLMM	
			smll	lrg	smll	lrg	smll	lrg	smll	lrg	smll	lrg	smll	lrg	smll	lrg
A	0.3	.5	5.7	5.7			5.7	5.2	5.5	5.5	11.0	14.8	5.3	5.6	7.5	6.9
		.8	5.4	5.0			5.4	5.3	5.3	4.9	10.9	10.2	5.2	5.0	5.1	4.9
		.9	5.2	5.0			5.2	5.3	5.2	5.0	9.5	6.2	5.4	5.4	5.5	2.1
	ne	.5	5.5	6.4			5.5	5.3	5.3	5.5	11.9	15.5	5.2	5.2	8.8	7.6
		.8	5.4	5.3			5.4	5.2	5.3	5.0	10.5	8.9	5.2	5.2	3.0	2.9
		.9	5.2	5.1			5.2	5.3	5.1	4.8	9.5	6.2	5.6	4.8	9.6	4.5

effect model	corr	p	parametric		param/HF		multivariate		Puri & Sen		ATS		Koch		GLMM	
			smll	lrg	smll	lrg	smll	lrg	smll	lrg	smll	lrg	smll	lrg	smll	lrg
A (B)	0.3	.5	5.4	5.4			5.4	5.3	5.3	5.2	10.7	13.0	5.3	5.3	7.1	5.7
		.8	5.3	5.5			5.3	5.3	5.2	5.1	11.9	10.9	5.3	5.1	5.8	6.3
		.9	5.3	5.4			5.3	5.3	5.3	5.2	11.2	8.4	5.6	5.2	3.6	3.2
	ne	.5	5.8	5.8			5.8	5.7	5.5	5.5	12.5	14.5	5.6	5.5	9.3	8.8
		.8	5.0	5.3			5.0	5.3	4.8	5.0	11.0	9.0	5.5	5.2	3.3	6.8
		.9	5.0	4.8			5.0	5.1	5.0	4.9	8.9	7.0	5.2	5.2	4.0	2.0
A (AB)	0.3	.5	5.6	5.8			5.6	5.5	5.2	5.2	11.3	14.2	5.5	5.4	3.7	3.9
		.8	5.3	5.5			5.3	5.5	5.2	5.2	11.5	10.3	5.2	5.3	7.3	3.4
		.9	5.1	5.3			5.1	5.3	5.0	5.0	9.6	7.5	5.4	5.4	7.3	5.6
	ne	.5	5.6	5.6			5.6	5.3	5.0	5.2	12.4	14.9	5.2	5.4	6.0	3.3
		.8	5.4	5.3			5.4	5.2	5.3	5.0	10.2	9.9	6.2	6.6	6.0	2.8
		.9	5.2	5.4			5.2	5.2	5.1	5.2	8.1	8.0	6.2	7.1	4.0	2.9
B	0.3	.5	5.3	5.3	5.4	5.4	4.9	5.3	4.9	5.2	5.8	5.2	4.6	4.9	6.3	6.4
		.8	4.9	5.5	4.7	5.4	4.8	5.9	4.8	5.4	4.8	5.4	4.4	5.8	4.2	4.4
		.9	4.9	5.1	4.4	4.7	5.2	5.6	4.8	4.9	4.1	4.6	4.7	5.1	5.8	2.9
	ne	.5	5.7	6.7	5.1	5.7	4.8	5.8	5.5	6.7	5.3	5.7	4.5	5.5	5.6	6.9
		.8	5.4	6.4	4.4	5.5	4.6	5.3	5.2	6.3	4.4	5.5	4.4	5.1	5.9	3.9
		.9	5.4	5.6	4.2	4.5	3.6	4.6	5.1	5.4	4.1	4.3	3.3	4.4	9.9	7.4
B (A)	0.3	.5	5.2	5.7	5.2	5.6	5.3	5.5	5.2	5.5	5.2	5.5	5.0	5.1	6.3	5.6
		.8	5.5	5.4	5.2	5.2	5.7	5.9	5.2	5.3	5.2	5.2	5.1	5.1	5.7	5.2
		.9	5.5	5.6	5.1	5.1	6.2	5.9	5.4	5.4	5.0	4.8	5.2	4.8	4.9	3.6
	ne	.5	5.9	6.6	5.2	5.6	5.5	5.5	5.7	6.1	5.3	5.4	5.2	5.3	4.8	5.5
		.8	6.2	6.2	5.5	5.5	5.3	5.5	6.0	6.0	5.6	5.5	4.6	4.8	7.7	5.8
		.9	5.5	6.4	4.2	5.2	4.8	5.6	5.2	6.1	4.3	4.9	3.6	5.0	9.8	6.3
B (AB)	0.3	.5	5.3	5.9	5.3	5.9	5.5	5.4	4.1	4.8	28.0	5.6	4.1	4.2	4.3	3.6
		.8	5.7	5.4	5.6	5.0	5.8	5.6	4.7	4.8	23.6	5.0	4.7	4.3	5.1	4.9
		.9	5.3	5.5	5.0	4.9	5.7	6.3	4.5	4.8	17.9	4.9	4.7	5.0	5.6	13.9
	ne	.5	5.6	7.3	5.0	6.2	4.8	5.1	4.5	6.2	24.1	6.0	3.7	3.9	3.6	4.9
		.8	5.6	6.4	4.6	5.1	4.7	5.1	4.3	5.6	17.3	5.1	3.8	4.3	7.2	4.2
		.9	5.6	6.4	4.6	5.1	4.3	5.6	4.8	5.8	10.4	5.1	3.5	4.8	10.0	8.1
AB	0.3	.5	5.7	5.8	5.6	5.8	5.5	5.3	5.6	5.8	5.7	5.7	5.2	5.2	7.0	8.7
		.8	5.4	5.8	5.2	5.5	5.2	5.4	5.3	5.4	5.2	4.6	5.1	5.0	4.8	3.9
		.9	5.5	6.8	5.2	5.0	5.0	5.2	5.0	5.4	5.2	4.4	4.5	4.6	5.9	3.0
	ne	.5	5.2	7.2	4.5	6.1	4.9	5.8	5.1	7.0	4.8	4.9	4.6	5.3	5.7	8.3
		.8	6.1	6.7	5.4	5.5	5.3	5.3	6.1	6.7	5.2	4.4	5.0	4.8	6.4	5.9
		.9	6.0	8.6	4.7	5.1	4.4	4.7	5.8	6.7	4.3	4.0	4.1	5.4	9.7	5.9
AB(A)	0.3	.5	5.1	5.4	5.2	5.3	5.3	5.6	5.1	5.2	5.2	5.0	4.9	5.1	7.0	7.2
		.8	5.6	7.1	5.3	6.6	5.9	5.4	5.2	6.5	5.3	4.7	6.4	6.7	1.8	1.7
		.9	5.7	8.1	5.4	7.6	6.0	5.7	5.4	7.7	5.4	4.7	6.5	7.9	2.2	0.9
	ne	.5	5.7	6.8	4.9	5.9	5.1	5.5	5.5	6.7	5.1	5.4	4.8	5.0	5.6	7.9
		.8	6.2	8.4	5.3	6.4	5.9	5.7	5.9	7.5	5.3	5.0	5.9	6.7	2.6	2.3
		.9	5.6	10.8	4.4	7.7	4.4	5.2	5.2	9.2	4.6	4.5	4.7	7.8	5.5	2.7
AB(B)	0.3	.5	5.5	5.3	5.6	5.4	5.4	5.7	2.8	2.2	5.6	5.1	5.1	5.0	7.7	7.1
		.8	6.2	6.1	5.1	5.2	5.3	5.3	3.8	3.4	6.3	4.9	5.1	4.9	2.7	2.4
		.9	7.0	7.7	5.1	6.1	3.8	5.1	4.3	4.0	6.9	5.1	5.8	5.6	3.8	2.7
	ne	.5	5.8	6.4	5.2	5.3	5.0	5.3	3.8	3.7	5.4	5.0	4.8	4.6	6.9	8.4
		.8	6.6	7.8	5.4	5.6	4.7	4.9	5.0	4.9	5.6	4.8	4.4	4.9	6.4	5.3
		.9	6.6	8.2	4.7	6.0	4.0	4.9	5.3	6.3	5.2	4.5	4.1	4.9	6.7	5.2

Table 1: Maximum smoothed type I error rates for all methods in all situations (mixed designs). Effects in brackets indicate other nonnull effects in the model.
Above: for small and large designs with independent n_i and p_i ,
below: for negative pairing (n_i ~ p_i) and positive pairing (n_i | p_i) in large designs.

effect model	corr	p	parametric		param/HF		multivariate		Puri & Sen		ATS		Koch		GLMM	
			$n_i \sim p_i$	n_i | p_i	$n_i \sim p_i$	n_i | p_i	$n_i \sim p_i$	n_i | p_i	$n_i \sim p_i$	n_i | p_i	$n_i \sim p_i$	n_i | p_i	$n_i \sim p_i$	n_i | p_i	$n_i \sim p_i$	n_i | p_i
B(A)	0.3	.5	5.2	5.4	5.2	5.4	5.4	5.5	5.3	5.3	4.9	5.1	4.8	5.3	4.9	5.4
		.8	5.9	5.3	5.7	5.2	6.0	5.5	5.7	5.5	5.3	4.8	5.7	5.2	3.7	4.1
		.9	5.5	5.1	5.2	4.8	5.7	5.0	5.3	5.1	4.4	4.6	5.5	4.8	2.7	3.1
	ne	.5	6.1	5.8	5.1	5.0	5.0	4.8	6.3	5.8	5.3	4.8	5.0	4.6	4.2	4.5
		.8	6.7	6.2	5.6	5.2	5.4	5.4	6.5	6.2	5.2	4.8	5.0	5.1	4.6	3.8
		.9	6.4	5.9	5.2	4.9	5.2	5.5	6.2	5.8	4.8	4.2	4.9	5.3	3.0	4.3
AB(A)	0.3	.5	6.9	3.9	6.9	3.9	6.7	4.0	5.3	4.7	4.7	4.6	6.7	3.7	5.2	4.8
		.8	16.1	2.3	15.4	2.3	11.6	2.4	15.4	2.2	4.7	3.6	15.3	2.2	3.0	1.1
		.9	20.7	1.7	19.8	1.6	14.4	1.6	20.3	1.6	4.0	3.5	21.1	1.5	2.2	0.3
	ne	.5	8.3	5.0	6.7	4.1	6.3	4.2	6.5	6.0	4.6	4.3	6.8	3.5	5.0	4.7
		.8	17.9	2.9	14.5	2.2	12.0	2.4	16.4	2.8	4.7	3.5	15.9	1.8	4.9	0.9
		.9	24.4	2.1	19.2	1.6	15.1	1.5	21.9	2.1	4.3	3.3	21.9	3.3	3.0	0.6

8. References

Agresti, A. (2002): *Categorical data analysis. Vol. 2*. New York, NY, John Wiley & Sons.

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

Akritas, M.G. & Brunner, E. (1997b): A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference*, 61, pp 249-277.

Bathke, A.C., Schabenberger, O., Tobias, R.D. & Madden, L.V. (2009): Greenhouse–Geisser Adjustment and the ANOVA-Type Statistic: Cousins or Twins?. The American Statistician, 63:3, pp 239-246.

Beckman, M. & Stroup, W.W. (2003): Small Sample Power Characteristics of Generalized Mixed Model Procedures for Binary Repeated Measures Data Using SAS. *Annual Conference on Applied Statistics in Agriculture*, Kansas State University Libraries, New Prairie Press.

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.

Brunner, E., Dette, H. & Munk, A. (1997): Box-Type Approximations in Nonparametric Factorial Designs, *Journal of the American Statistical Association*, Vol. 92, No. 440, pp. 1494- 1502.

Brunner, E., Konietschke, F., Pauly, M. and Puri, M.L. (1999a): Rank-Based Procedures in Factorial Designs: Hypotheses about Nonparametric Treatment Effects, *Journal of the Royal Statistical Society,* Series B (Statistical Methodology) 79(5).

Brunner, E., Munzel, U. and Puri, M.L. (1999b): Rank-Score Tests in Factorial Designs with Repeated Measures*, Journal of Multivariate Analysis* 70, pp 286-317.

Carr, J.C. & Chi, E.M. (1992): Analysis of Variance for Repeated Measures Data: A Generalized Estimating Equations Approach. *Statistics in Medicine*, Vol 11, pp 1033-1040.

Chatterjee,S . K. and Sen, P. K. (1966): Non-parametric tests for the multivariate multisample

location problem, *S. N. Roy Memorial Volume*, edited by R.C. Bose, et. al.,

Cleary, P.D. & Angel, R. (1984): The Analysis of Relationships Involving Dichotomous Dependent Variables. *Journal of Health and Social Behavior*, Vol. 25, No. 3, pp. 334-348.

Cochran, W.G. (1950): The Comparison of Percentages in Matched Samples. Biometrika 37, pp 256-266.

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

D'Agostino, R.B. (1972): Relation Between the Chi-Squared and ANOVA Tests for Testing the Equality of k Independent Dichotomous Population. *The American Statistician*, Vol. 26, No. 3, pp. 30-32.

D'Agostino, R.B. (1971): A Second Look at Analysis of Variance on Dichotomous Data, *Journal of Educational Measurement*, Vol. 8, No. 4, pp. 327-333.

Draper, J. F. (1972): *A Monte Carlo Investigation of the Analysis of Variance Applied to non-independent Bernoulli Variables*. Annual meeting of the American Educational Research Association, Chicago,Illinois

Emrich L.J. , Piedmonte M.R. (1992): On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation*, 41, 19-29 .

Ernst, M.D. & Kepner, J.l. (1993) A monte carlo study of rank tests for repeated measures designs, *Communications in Statistics - Simulation and Computation*, 22:3, pp 671-678,

Fan, C., Zhang, D. & Zhang, C.H. (2013): A comparison of bias-corrected covariance estimators for generalized estimating equations. *Journal of Biopharmaceutical Statistics* 23, pp 1172–1187.

Fan, C. & Zhang, D. (2014): Robust small sample inference for generalised estimating equations: An application of the Anova-type test. *Australian & New Zealand Journal of Statistics*, 56(3), pp 237–255.

Fay, M. P., Graubard, B. I. (2001): Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 57, pp 1198–1206.

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.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage Publications.

Fox, J. & Weisberg, S. (2015): *Mixed-Effects Models in R, an Appendix to An R Companion to Applied Regression*. SAGE Publications, Los Angeles.

Gosho M, Sato Y, Takeuchi H. (2014): Robust covariance estimator for small-sample adjustment in the generalized estimating equations: A simulation study. *Science Journal of Applied Mathematics and Statistics*, 2(1), pp 20–25.

Guerin, L., Stroup, W.W. (2000): A Simulation Study to Evaluate PROC MIXED Analysis of Repeated Measures Data. *Annual Conference on Applied Statistics in Agriculture.* URL: http://newprairiepress.org/agstatconference/2000/proceedings/15

Harville, D.A. (1977): Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, Vol. 72, No. 358, pp. 320-338 .

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992): Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17, pp 315-339.

Harwell, M.R. & Serlin, R.C. (1989): A Nonparametric Test Statistic for the General Linear Model. *Journal of Educational Statistics*, Vol. 14, No. 4, pp 351-371.

Harwell, M.R. & Serlin, R.C. (1994): A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances, *Computational Statistics & Data Analysis,* 17, pp 35-49.

Hsu, T.Chi and Feldt, L.S. (1969): The Effect of Limitations on the Number of Criterion Score Values on the Significance Level of the F-Test, *American Educational Research Journal*, Vol. 6, No. 4 (Nov., 1969), pp. 515-527.

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

Jaeger, T.F. (2008) Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models, J*ournal of Memory and Language*, 59(4): pp 434–446.

Kauermann, G., Carroll, R.J. (2001): A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 96:pp 1387–1396.

Kenward, M.G and Jones, B (1992: Alternative approaches to the analysis of binary and categorical repeated measurements. *Journal of Biopharmaceutical Statistics,* 2(2), pp 137-170.

Koch, G.G. (1969): Some Aspects of the Statistical Analysis of Split Plot Experiments in completely Randomized Designs, *Journal of the American Statistical Association*, Vol 64, No 326, pp 485-504.

Koch, G.G. (1970): The Use of Non-Parametric Methods in the Statistical Analysis of a Complex Split Plot Experiment. *Biometrics*, Vol. 26, No. 1, pp. 105-128.

Koch, G.G, Landis, J.R *et al* (1977): A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics,* 33, pp 133-158.

Konietschke, F., Bathke, A.C., Hothorn, L.A., Brunner, E. (2010): Testing and estimation of purely nonparametric effects in repeated measures designs, *Computational Statistics & Data Analysis,* 54(8):1895-1905.

Leisch, F., Weingessel, A., Hornik, K. (1998): On the Generation of Correlated Artificial Binary Data. Working Paper Series, Vienna University of Economicsand Business Administration, URL: http://epub.wu.ac.at/286/1/document.pdf.

Li, P. & Redden, D.T. (2015): Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*,

https://doi.org/10.1186/s12874-015-0026-x

Liang, K.Y. & Zeger S.L. (1986): A Comparison of Two Bias-Corrected Covariance Estimators for Generalized Estimating Equations. *Biometrika* 73,pp 13–22.

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 nicht-parametrischen 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 aligned rank transform and discrete variables: A warning, *Communications in Statistics - Simulation and Computation*,
DOI: 10.1080/03610918.2016.1217014

Luepsen, H. (2017):Comparison of nonparametric analysis of variance methods: A Vote for van der Waerden. Communications in Statistics - Simulation and Computation, Volume 30, pp 1-30, DOI: 10.1080/03610918.2017.1353613

Lunney, G.H. (1979): Using Analysis of Variance with a Dichotomous Dependent Variable: An Empirical Study. *Journal of Educational Measurement,* Vol. 7, No. 4, pp. 263-269.

Ma, Y., Mazumdar, M., Memtsoudis, S.G. (2012): Beyond Repeated measures ANOVA: advanced statistical methods for the analysis of longitudinal data in anesthesia research, *Reg Anesth Pain Med*, 37(1): pp 99–105.

Malhotra, N.K. (1983): A Comparison of the Predictive Validity of Procedures for Analyzing Binary Data, *Journal of Business & Economic Statistics,* Vol. 1, No. 4, pp. 326-336.

Mancl, L. A., DeRouen, T. A. (2001): A covariance estimator for GEE with improved smallsample properties. *Biometrics* 57, pp 126–134.

Mandeville, G.K. (1972): Comparison of Three Methods of Analyzing Dichotomous Data in a Randomized Flock Design, Distributed by ERIC Clearinghouse.

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

McNeish, D. & Stapleton, L.M. (2016): Modeling Clustered Data with Very Few Clusters, *Multivariate Behavioral Research*, 51 (4), pp 495-518.

McNeish, D. & Harring, J.R. (2017): Clustered data with small sample sizes: Comparing the performance of model-based and designbased approaches, *Communications in Statistics - Simulation and Computation*, 46 (2), pp 855-869.

Morel, J.G., Bokossa, M.C., Neerchal, N.K. (2003): Small sample correction for the variance of GEE estimators. *Biometrical Journal* 45, pp 395–409.

Noguchi, K., Gel, Y.R., Brunner, E., and Konietschke, F. (2012). nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments. *Journal of Statistical Software* , 50 (12), pp 1-23.

Oberfeld, D. & Franke, T. (2012): Evaluating the. robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavioural Research*, 45: pp 792–812.

Pan, W. (2001): On the Robust Variance Estimator in Generalized Estimation Equations. *Biometrika* 88, No 3, pp 901-906.

Pan, W. & Wall, M.M. (2001): Small Sample Adjustments in Using the Sandwich Variance Estimator in Generalized Estimating Equations. *Statistics in Medicine*, Volume 21, Issue 10, pp 1429–1441.

Pan, W. & Connett, J.E. (2002): Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica*, Vol 12, No 2, pp 475-490.

Peng C.Y.J., Lee K.L. , Ingersoll G.M. (2002): An introduction to logistic regression analysis and reporting, *The Journal of Educational Research*, Vol 96, No 1, pp 3-14.

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

Pohlmann, John T. & Leitner, Dennis W. (2003): A comparison of ordinary least squares and logistic regression. *The Ohio Journal of Science.* 103.5, pp118-125.

Prentice, R.L. & Zhao, L.P. (1991): Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. *Biometrics*, Vol. 47, No. 3, pp 825-839.

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

Qu, Y., Piedmonte, M.R. & Williams, G.V. (1994): Small Sample Validity of Latent Variable Models for Correlated Binary Data. *Communications in Statistics - Simulation and Computation*, Vol 23, No 1. pp 243-269.

Richter, S.J. and Payton, M. (1999). Nearly exact tests in fact orial experiments using the aligned rank transform. *Journal of Applied Statistics*, Volume 26, Issue 2, pp. 203-217.

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

Song, X.Y. & Lee, S.Y. (2006): Model comparison of generalized linear mixed models. *Statistics in Medicine*, 25, pp 1685–1698.

Stiger, R.T., Kosinski, A.S. , Barnhart, H.X. & Kleinbaum, D.G. (1998) Anova for repeated ordinal data with small sample size? A comparison of anova, manova, wls and gee methods by simulation, *Communications in Statistics - Simulation and Computation*, 27:2, pp 357-375.

Swafford, M. (1980): Three Parametric Techniques for Contingency Table Analysis: A Nontechnical Commentary. *American Sociological Review*, 45, pp 664-690.

Tandon, P.K. & Moeschberger, M.L. (1989) Comparison of Nonparametric and Parametric Methods in Repeated Measures Designs - A Simulation Study, *Communications in Statistics - Simulation and Computation*, 18:2, pp 777-792.

Tansey, R., White, M., Long, R.G., Smith, M. (1996): A Comparison of Loglinear Modeling and Logistic Regression in Management Research. *Journal of Management*, 22, No 2, pp 339-358.

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.

Tuerlinckx, F., Rijmen, F., Verbeke, G. & De Boeck, P. (2006): Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59, pp 225–255.

Wang M. & Long Q. (2011): Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Statistics in medicine*, 30(11), pp 1278–1291.

Wang, M., Kong, L., Zheng, L. & Zhang, L. (2016): Covariance estimators for Generalized Estimating Equations (GEE) in longitudinal analysis with small samples. *Statistics in* Winer, B.J., Brown, D.R. & Michels, K.M. (1991): *Statistical Principles in Expertimental Design*, McGraw-Hill, New York.

Zhang, H., N. Lu, C. Feng, S. Thurston, Y. Xia, L. Zhu, and X. Tu (2011): On Fitting Generalized Linear Mixed-effects Models for Binary Responses using Different Statistical Packages. *Statistics in Medicine*, 30(20), pp 2562–2572.

Ziegler, A., Kastner, Ch., Blettner, M. (1998): The Generalised Estimating Equations: An Annotated Bibliography. Biometrical Journal 40 (2), pp 115-139.