Journal of Statistical Software

July 2017, Volume 79, Code Snippet 1. *doi: 10.18637/jss.v079.c01*

GFD: An R Package for the Analysis of General Factorial Designs

Sarah Friedrich
Ulm University

Frank Konietschke
UT Dallas

Markus Pauly
Ulm University

Abstract

Factorial designs are widely used tools for modeling statistical experiments in all kinds of disciplines, e.g., biology, psychology, econometrics and medicine. For testing null hypotheses in this framework, ANOVA methods are widely used. However, the corresponding F tests are only valid for normally distributed data with equal variances, two assumptions which are often not met in practice. The R package **GFD** provides an implementation of the Wald-type statistic (WTS), the ANOVA-type statistic (ATS) and a studentized permutation version of the WTS. Both the WTS and the permuted WTS do not require normally distributed data or variance homogeneity, whereas the ATS assumes normality. All methods are available for general crossed or nested designs and all main and interaction effects can be plotted. Additionally, the package is equipped with an optional graphical user interface to facilitate application for a wide range of users. We illustrate the implemented methods for a range of different designs.

Keywords: factorial designs, non-normal data, heteroscedasticity, permutation, R, GUI.

1. Introduction

Originated in the agricultural sciences factorial designs are widely used tools for modeling statistical experiments in a variety of disciplines, e.g., biology, econometrics, medicine, ecology or psychology. For testing null hypotheses formulated in terms of means, analysis-of-variance (ANOVA) methods are well known, and preferred for making statistical inference. ANOVA methods are implemented in R within the function `aov` in the R package **stats** (R Core Team 2017). The `anova` function in this package as well as `Anova` in the **car** package (Fox and Weisberg 2011) provide clearly arranged ANOVA tables for fitted models. The corresponding F tests, however, are only valid under the assumption of normally distributed errors and equal variances across the different treatment groups. These assumptions are hard to verify in practice and often not met. A violation usually inflates the type-I or -II errors of the F statistics.

The accuracy of the F tests depends on the actual data distributions, sample size allocations, and the degree of variance heteroscedasticity. For normally distributed errors, several procedures for heteroscedastic data have been proposed, e.g., the generalized Welch-James test (Johansen 1980), the approximate degrees of freedom test (Zhang 2012) or the ANOVA-type test proposed by Brunner, Dette, and Munk (1997), see also Bathke, Schabenberger, Tobias, and Madden (2009). These tests control the type-1 error level in heteroscedastic designs quite accurately, but are in general not asymptotically exact for non-normal data. In comparison to that, the Wald-type statistic, see Equation 2 below, is asymptotically exact in general factorial designs without assuming variance homogeneity or normally distributed error terms. It is well known, however, that the Wald-type statistic requires large sample sizes to control the pre-assigned type-I error, see e.g., Vallejo, Fernández, and Livacic-Rojas (2010). Its small sample behavior may be improved by applying an adequate permutation procedure, see Pauly, Brunner, and Konietschke (2015) for the theoretical background. The only comparable test included in the R function `oneway.test` is the Welch (1951) test for heteroscedastic one-way layouts. Furthermore, an ANOVA-type test based on ranks is also implemented in the R package **asbio** (Aho 2017) within the functions `BDM` and `BDM.2way` for nonparametric one- and two-way layouts, respectively.

For a user friendly application of these rather robust methods in statistical data sciences, the R package **GFD** has been developed. The use of the main function `GFD` as well as its output are very similar to the `aov` function from the R package **stats** or the `Anova` function from the R package **car** (Fox and Weisberg 2011). Its application provides a descriptive overview of the data as well as the complete ANOVA-tables according to the `formula` input, which allows the modeling of arbitrary high-way layouts. Hereby the Wald-type statistic, a permuted version thereof as well as the ANOVA-type statistic for these general factorial designs are implemented. Both the Wald-type statistic as well as the permutation test neither assume normality nor homogeneous variances, while the ANOVA-type statistic assumes normality. Furthermore, all main and interaction effects can be plotted along with $(1-\alpha)$ confidence intervals. In addition, the package is equipped with a graphical user interface (GUI) to facilitate application for a wide audience of statisticians, practitioners, and educational purposes. The package is freely available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=GFD`.

The paper is organized as follows: In Section 2 we describe the statistical model and the tests used in this setting. In Section 3 we provide various examples for different settings which are statistically evaluated with the R package **GFD**. Finally, we discuss the results in Section 4 and provide an outlook to future work.

Throughout the paper we use the following notation: We denote by $\boldsymbol{P}_a = \boldsymbol{I}_a - \frac{1}{a}\boldsymbol{J}_a$ the a-dimensional centering matrix, $\boldsymbol{I}_a$ is the a-dimensional unit matrix and $\boldsymbol{J}_a$ denotes the $a \times a$ matrix of 1's, i.e., $\boldsymbol{J}_a = \mathbf{1}_a \mathbf{1}_a^\top$, where $\mathbf{1}_a = (1, \ldots, 1)^\top$ is the a-dimensional column vector of 1's.

2. Statistical model and inference methods

In order to cover different factorial designs, we consider the following general linear model

$$Y_{ik} = \mu_i + \varepsilon_{ik}, \tag{1}$$

where $k = 1, \ldots, n_i$ is the experimental unit within class $i = 1, \ldots, a$. Note that different

sample sizes n_i are admitted. For each fixed i the error terms ε_{ik} are independent and identically distributed with $\mathsf{E}(\varepsilon_{i1}) = 0$ and $\mathsf{VAR}(\varepsilon_{i1}) = \sigma_i^2 > 0$. Note that we neither assume normality of the error terms nor variance homoscedasticity. In this setting, a higher way factorial structure with crossed or nested factors can be achieved by splitting up the index i into sub-indices $i_1, i_2, \ldots, i_p$. In our notation, the components $i = 1, \ldots, a$ can be considered as a lexicographic order of the factor level combinations.

In this framework we like to test general linear null hypotheses

$$H_0^{\mu} : \boldsymbol{H}\boldsymbol{\mu} = \mathbf{0}$$

about the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_a)^\top$. Here $\boldsymbol{H}$ denotes an adequate hypothesis contrast matrix of interest.

Let $\overline{\boldsymbol{Y}}_{\cdot} = (\overline{Y}_{1\cdot}, \ldots, \overline{Y}_{a\cdot})^\top$ denote the vector of group means and let $\boldsymbol{V}_N = \mathsf{COV}(\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}) = \mathrm{diag}(\frac{N}{n_i}\sigma_i^2 : i = 1, \ldots, a)$ denote the covariance matrix of $\sqrt{N}\,\overline{\boldsymbol{Y}}_{\cdot}$. Then $\boldsymbol{V}_N$ is consistently estimated by $\widehat{\boldsymbol{V}}_N = \mathrm{diag}(\frac{N}{n_i}\widehat{\sigma}_i^2)$, where $\widehat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}(Y_{ik} - \overline{Y}_{i\cdot})^2$ denotes the empirical variance of the sample $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$.

In order to test the null hypotheses formulated above in this general framework, we consider two generalizations of the two-sample Welch t statistic: The Wald-type statistic (WTS) as discussed, e.g., in Pauly *et al.* (2015), and the ANOVA-type statistic (ATS) from Brunner *et al.* (1997). The WTS is given by

$$Q_N = N\,\overline{\boldsymbol{Y}}_{\cdot}^{\top}\boldsymbol{H}^{\top}(\boldsymbol{H}\widehat{\boldsymbol{V}}_N\boldsymbol{H}^{\top})^{+}\boldsymbol{H}\,\overline{\boldsymbol{Y}}_{\cdot}. \tag{2}$$

Here, $\boldsymbol{M}^{+}$ denotes the Moore-Penrose inverse of a matrix $\boldsymbol{M}$. It is well known that under rather weak assumptions the WTS has asymptotically a central χ_f^2 distribution with $f = \mathrm{rank}(\boldsymbol{H})$ degrees of freedom under $H_0^{\mu} : \boldsymbol{H}\boldsymbol{\mu} = \mathbf{0}$. However, the WTS requires large sample sizes to get a satisfactory approximation by using the quantiles of the limiting χ^2 distribution (Akritas, Arnold, and Brunner 1997; Akritas and Brunner 1997; Vallejo *et al.* 2010; Pauly *et al.* 2015).

A second generalization of the two-sample Welch statistic is the ANOVA-type statistic (ATS) defined as

$$A_N = \frac{N}{\mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{V}}_N)}\overline{\boldsymbol{Y}}_{\cdot}^{\top}\boldsymbol{T}\,\overline{\boldsymbol{Y}}_{\cdot},$$

where $\boldsymbol{T} = \boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{H}^{\top})^{-}\boldsymbol{H}$. Following Brunner *et al.* (1997) the distribution of the ATS can be approximated by an $F(\hat{f}, \hat{f}_0)$-distribution such that the first two moments coincide, i.e., by choosing

$$\hat{f} = \mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{V}}_N)^2/\mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{V}}_N\boldsymbol{T}\widehat{\boldsymbol{V}}_N)$$

and

$$\hat{f}_0 = \mathrm{tr}(\boldsymbol{T}\widehat{\boldsymbol{V}}_N)^2/\mathrm{tr}(\boldsymbol{D}^2\widehat{\boldsymbol{V}}_N^2\boldsymbol{\Lambda}).$$

Here $\boldsymbol{D}$ denotes the matrix of diagonal elements of $\boldsymbol{T}$ and $\boldsymbol{\Lambda} = \mathrm{diag}((n_1 - 1)^{-1}, \ldots, (n_a - 1)^{-1})$ (Brunner *et al.* 1997; Brunner and Puri 2001). Note that in the two-sample case this approximation coincides with the Satterthwaite-t-approximation. However, the ATS is in general asymptotically exact only for normally distributed error terms.

Another possibility is to improve the small sample behavior of the WTS by applying a permutation procedure (Pauly *et al.* 2015). To describe this procedure in detail, let $\boldsymbol{Y}^{\pi} =$

$\pi(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_a)^\top$ denote a fixed but arbitrary permutation of $\boldsymbol{Y}$, i.e., $\pi \in \mathcal{S}_N$. Furthermore, let $\overline{\boldsymbol{Y}}_\cdot^\pi = (\overline{Y}_{1\cdot}^\pi, \ldots, \overline{Y}_{a\cdot}^\pi)^\top$ denote the vector of means and $\widehat{\boldsymbol{V}}_N^\pi = \operatorname{diag}\left(\frac{N}{n_i}(\widehat{\sigma}_i^\pi)^2 : i = 1, \ldots, a\right)$ the diagonal matrix of empirical variances $(\widehat{\sigma}_i^\pi)^2$ under this permutation. Then, the permuted Wald-type statistic (WTPS) is given by

$$Q_N^\pi = N(\overline{\boldsymbol{Y}}_\cdot^\pi)^\top \boldsymbol{H}^\top (\boldsymbol{H}\widehat{\boldsymbol{V}}_N^\pi \boldsymbol{H}^\top)^+ \boldsymbol{H}\,\overline{\boldsymbol{Y}}_\cdot^\pi,$$

which is the WTS as defined in Equation 2 calculated with the permuted observations. Now, a permutation test is achieved by the following steps:

1. Fix the data $\boldsymbol{Y}$ and compute the WTS Q_N.
2. Permute the data randomly and obtain the value of Q_N^π. Safe this in A_1.
3. Repeat Step 2 J (say $J = 10,000$) times and obtain the values $A_1, \ldots, A_J$.
4. Compute the p value by the (approximative) conditional permutation distribution (i.e., the empirical distribution of $A_1, \ldots, A_J$) as

$$p \text{ value} = \frac{1}{J}\sum_{j=1}^{J} \mathcal{I}(Q_N \geq A_j).$$

Instead of computing the p value for making statistical inference, the original WTS Q_N can be compared with the $(1-\alpha)$ quantile of the conditional distribution of Q_N^π given the data $\boldsymbol{Y}$, i.e., the empirical quantile of $A_1, \ldots, A_J$. Pauly *et al.* (2015) have shown that this algorithm yields a valid permutation approach and consistent level α test, i.e., the conditional distribution of the WTPS always approximates the null distribution of Q_N. The test controls the preassigned level α under the null hypothesis and is even finitely exact if the pooled data is exchangeable under the hypothesis. Note that in the special case of a one-way layout the WTPS reduces to the permutation test for means of Chung and Romano (2013). The default value for the number of permutation runs in the R package **GFD** is nperm $= J = 10,000$.

For practical recommendations we briefly summarize the main properties of the three considered tests from Pauly *et al.* (2015): Mathematically, only the WTS and WTPS provide valid asymptotic procedures for general factorial designs. Nevertheless, simulation studies demonstrate that the ATS controls the α level for finite samples rather satisfactory. In case of non-normal data, however, the test tends to be conservative, which leads to loss of power. The WTS, in contrast, is quite liberal for small to moderate sample sizes. The WTPS is a rather accurate procedure even for non-normal data. When data is very skewed and heteroscedastic, the test tends to be liberal and to over-reject the hypothesis, in particular when the larger sample has the smaller variance (so called negative pairing). Its liberality is, however, not as pronounced as for the WTS.

Note that in comparison the **coin** package (Hothorn, Hornik, van de Wiel, and Zeileis 2008), which contains permutation tests for two- and multiple-sample problems, does not, e.g., handle heteroscedastic shift models. In our more general situation we allow for different variances and/or different distributions among the different groups. Furthermore, the Welch test from the function `oneway.test` is also only an approximation for normally distributed models that is known to perform worse than the ATS and the WTPS, see e.g., Vallejo *et al.* (2010) and

Pauly *et al.* (2015). Remark further, that the ANOVA-type tests from the R package **asbio** (Aho 2017) are based on ranks and test different null hypotheses formulated in terms of distribution functions instead of means.

For the calculation of the confidence intervals, we have used the corresponding quantiles of the t distribution.

2.1. Two-sample tests

A special case of model (1) is the heteroscedastic two-sample case, i.e., $a = 2$. This results in the extended Behrens-Fisher model

$$Y_{ik} = \mu_i + \varepsilon_{ik}, \; i = 1, 2; \; k = 1, \ldots, n_i,$$

which is usually analyzed using a Welch's t test in the statistic

$$T_N = \frac{\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}}{\sqrt{\widehat{\sigma}_1^2/n_1 + \widehat{\sigma}_2^2/n_2}}. \tag{3}$$

Its distribution is approximated by a t_ν distribution with estimated Satterthwaite-Welch degree of freedom ν to account for variance heterogeneity. Another possibility to approximate the distribution of T_N as defined in Equation 3 is to employ the studentized permutation distribution of T_N, and to carry out the test as a permutation test as proposed by Janssen (1997, 2005).

Note that the Wald-type statistic Q_N, as well as the ATS A_N are the square of T_N in the two-sample case. Furthermore, both the statistics Q_N and A_N are identical in this setup; and the second degree of freedom $\hat{f}_0$ of the ATS is identical to the Satterthwaite-Welch degree of freedom. The first degree of freedom $\hat{f}$ is equal to 1, by definition. Thus the ATS test is essentially Welch's t test and the WTPS test is in fact Janssen's permutation test.

3. Examples

In this section, we provide examples demonstrating how different factorial designs can be analyzed using the **GFD** package. The function `GFD` returns an object of class '`GFD`' from which the user may obtain plots and summaries of the results using `plot()`, `print()` and `summary()` methods, respectively. Here, `print()` returns a short summary of the results, i.e., the values of the test statistics along with degrees of freedom and corresponding p values whereas `summary()` also displays some descriptive statistics such as the means and variances for the different factor level combinations. Plotting is based on **plotrix** (Lemon 2006). For two- and higher-way layouts, the factors for plotting can be additionally specified in the `plot` call, see the examples below.

```
GFD(formula, data = NULL, nperm = 10000, alpha = 0.05)
```

Note that the test statistics for the main effects considered in Section 2 are not changed by whether or not an additional interaction term is specified in `formula` since the tests are determined by the choice of the hypothesis matrix $\boldsymbol{H}$. Only crossed and hierarchical (nested) designs are implemented – a mixture of both is up to date not available.

Figure 1: The GUI for tests in general factorial designs: The user can specify the data location, the formula, the number of permutations and the significance level α. One can additionally choose to plot the results.

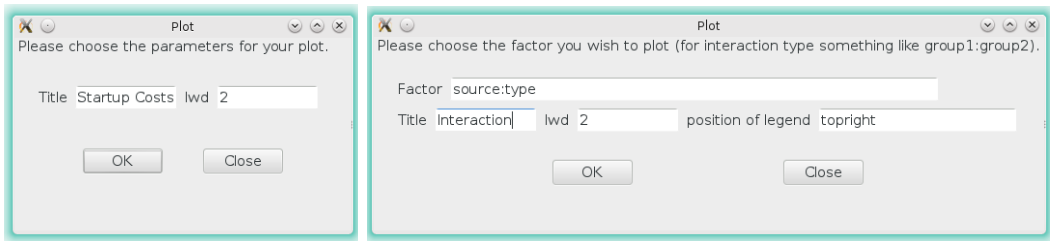

Figure 2: Graphical user interfaces for plotting: The left GUI is for the one-way layout (no choice of factors possible), the other one is for a higher-way layout. An example for plotting interactions is given in the right panel.

Furthermore, the **GFD** package is equipped with an optional GUI, based on **RGtk2** (Lawrence and Temple Lang 2010), which will be explained in detail in the next section.

3.1. Graphical user interface

The GUI is started in R with the command `calculateGUI()`. Note that the GUI depends on **RGtk2** and will only work if **RGtk2** is installed. The user can specify the data location (either directly or via the "load data" button), the formula, the number of permutations and the significance level α, see Figure 1. Additionally, one can specify whether or not headers are included in the data file, and which separator and character symbols are used for decimals in the data file. The GUI also provides a plotting option, which generates a new window for specifying the factors to be plotted (in higher-way layouts) along with a few plotting parameters, see Figure 2. Note that four- and higher way interactions cannot be plotted due to the increasing complexity of the plots.

```
R> library("GFD")
R> calculateGUI()
```

3.2. Two-sample tests

As an example of a two-sample problem we consider a subset of the `weightgain` data set (Hand, Daly, McConway, Lunn, and Ostrowski 1993) from the **HSAUR** package (Everitt and Hothorn 2017). The data contains information on the weight gain (in grams) of rats which

were randomized to one of four diets, distinguished by the amount of protein (high and low) and the source of protein (beef and cereal). For our purposes, we first restrict our analysis to the high protein group.

```
R> library("GFD")
R> data("weightgain", package = "HSAUR")
R> weightgain2 <- subset(weightgain, type == "High")
R> set.seed(123)
R> two_sample <- GFD(weightgain ~ source, data = weightgain2,
+    nperm = 10000, alpha = 0.05)
R> plot(two_sample, main = "Two-sample test", cex.axis = 1.5,
+    cex.lab = 1.5, cex.main = 1.5, lwd = 2)
R> two_sample

Call:
weightgain ~ source

Wald-Type Statistic (WTS):
Test statistic             df        p-value   p-value WTPS
    4.37169244     1.00000000     0.03654068     0.05580000

ANOVA-Type Statistic (ATS):
Test statistic            df1            df2        p-value
    4.37169244     1.00000000    17.99896078     0.05099558
```

Note that the results are identical with those using the `t.test` function:

```
R> t.test(weightgain ~ source, data = weightgain2)

Welch Two Sample t-test

data:  weightgain by source
t = 2.0909, df = 17.999, p-value = 0.051
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.0679184 28.2679184
sample estimates:
  mean in group Beef mean in group Cereal
               100.0                 85.9
```

As mentioned in Section 2.1 the p values obtained using the ATS and the Satterthwaite-Welch t test are identical. A reason for the smaller p value obtained with the WTS may be given due to its more liberal behavior in case of small sample sizes ($n_1 = n_2 = 10$), see Vallejo *et al.* (2010) and Pauly *et al.* (2015).

The data may also be analyzed using the GUI, see Figure 3 for an example. The corresponding plot of the effect is given in Figure 4.

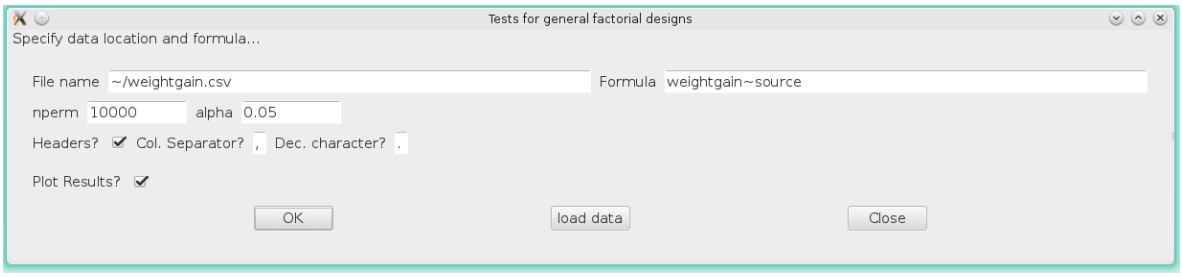

Figure 3: Graphical user interface with formula for the `weightgain` data set.

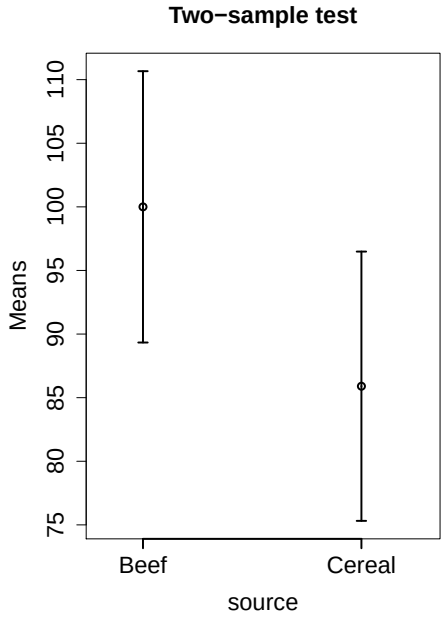

Figure 4: Mean weight gain for the two different sources of protein, beef and cereal, in the two-sample problem.

3.3. One-way layout

In a one-way layout,

$$Y_{ik} = \mu_i + \varepsilon_{ik},\ i = 1, \ldots, a;\ k = 1, \ldots, n_i,$$

we are interested in the effect of factor A, i.e., we wish to test the null hypothesis $H_0 : \{\mu_1 = \ldots = \mu_a\} = \{\boldsymbol{P}_a\boldsymbol{\mu} = \mathbf{0}\}$.

An example for such a model is the data set on startup costs of companies, which was selected from the Business Opportunities Handbook, see Cengage College (2008). The data represent business startup costs in thousands of dollars for five different kinds of shops.

```
R> library("GFD")
R> data("startup", package = "GFD")
R> set.seed(456)
R> model1 <- GFD(Costs ~ company, data = startup, nperm = 10000,
+    alpha = 0.05)
```

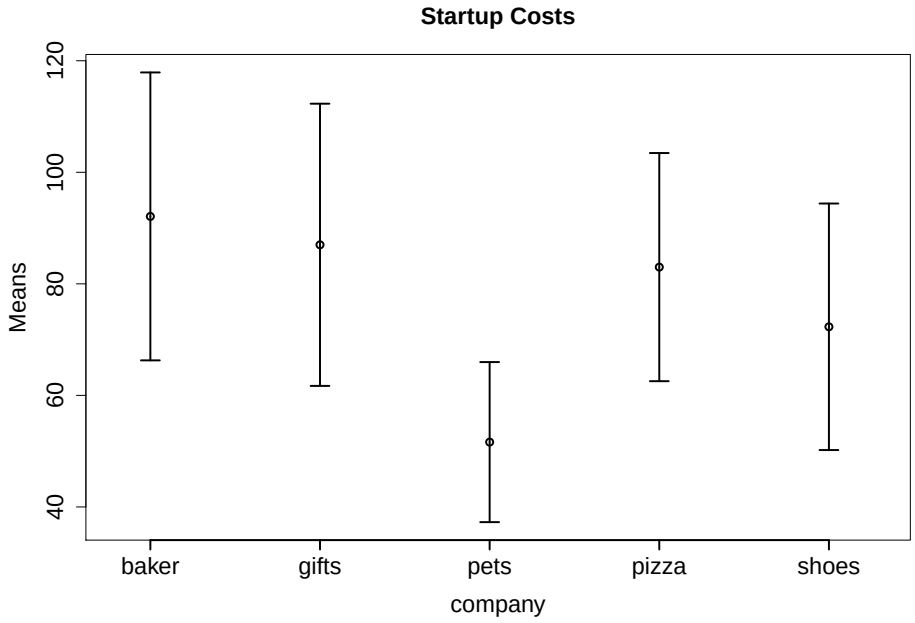

Figure 5: Mean startup costs for the five different companies in the `startup` data example.

```
R> summary(model1)
R> plot(model1, main = "Startup Costs", cex.axis = 1.5, cex.lab = 1.5,
+    cex.main = 1.5, lwd = 2)

Call:
Costs ~ company

Descriptive:
  company  n    Means Variances Lower 95 % CI Upper 95 % CI
1   baker 11 92.09091 1512.6909      66.28044     117.90138
2   gifts 10 87.00000 1289.1111      61.70193     112.29807
3    pets 16 51.62500  733.0500      37.27595      65.97405
4   pizza 13 83.00000 1165.1667      62.54732     103.45268
5   shoes 10 72.30000  983.7889      50.19995      94.40005

Wald-Type Statistic (WTS):
Test statistic             df        p-value   p-value WTPS
  15.037830399    4.000000000    0.004623394    0.024600000

ANOVA-Type Statistic (ATS):
Test statistic            df1            df2        p-value
    2.57248203     3.70623134    44.51042721     0.05456579
```

This example nicely demonstrates the liberal behavior of the WTS (p value = 0.0046) as well as the conservative behavior of the ATS (p value = 0.055). The WTPS, in contrast, is somewhere in between with a p value of 0.0246.

3.4. Two-way layout

In a two-way crossed design,

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

with $i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, n_{ij}$, one is interested in tests for the main effects of the factors A and B as well as for an interaction of the two, i.e.,

$$\begin{aligned} H_0(A) &: \{\alpha_i = \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot\cdot} = 0 \ \forall\ i = 1, \ldots, a\}, \\ H_0(B) &: \{\beta_j = \overline{\mu}_{\cdot j} - \overline{\mu}_{\cdot\cdot} = 0 \ \forall\ j = 1, \ldots, b\}, \\ H_0(AB) &: \{\gamma_{ij} = \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \overline{\mu}_{\cdot\cdot} = 0 \ \forall\ i = 1, \ldots, a,\ j = 1, \ldots, b\}, \end{aligned}$$

or formulated with suitable contrast matrices:

$$\begin{aligned} H_0(A) &: \{\boldsymbol{H}_A\boldsymbol{\mu} = \boldsymbol{P}_a \otimes \frac{1}{b}\mathbf{1}_b^\top \cdot \boldsymbol{\mu} = \mathbf{0}\}, \\ H_0(B) &: \{\boldsymbol{H}_B\boldsymbol{\mu} = \frac{1}{a}\mathbf{1}_a^\top \otimes \boldsymbol{P}_b \cdot \boldsymbol{\mu} = \mathbf{0}\}, \\ H_0(AB) &: \{\boldsymbol{H}_{AB}\boldsymbol{\mu} = \boldsymbol{P}_a \otimes \boldsymbol{P}_b \cdot \boldsymbol{\mu} = \mathbf{0}\}. \end{aligned}$$

We will again consider the `weightgain` data set from package **HSAUR**. This time, however, we are interested in analyzing both factors, i.e., amount and source of protein.

```
R> library("GFD")
R> data("weightgain", package = "HSAUR")
R> set.seed(789)
R> model2 <- GFD(weightgain ~ source * type, data = weightgain)
R> summary(model2)
R> plot(model2, factor = "source:type", main = "Interaction", xlab = "Type",
+    cex.axis = 1.5, cex.lab = 1.5, cex.main = 1.5)
R> plot(model2, factor = "source", main = "Mean weight gain",
+    xlab = "source", cex.axis = 1.5, cex.lab = 1.5, cex.main = 1.5)

Call:
weightgain ~ source * type

Descriptive:
  source type  n Means Variances Lower 95 % CI Upper 95 % CI
1   Beef High 10 100.0  229.1111      89.33489     110.66511
3   Beef  Low 10  79.2  192.8444      69.41534      88.98466
2 Cereal High 10  85.9  225.6556      75.31562      96.48438
4 Cereal  Low 10  83.9  246.7667      72.83158      94.96842

Wald-Type Statistic (WTS):
            Test statistic df    p-value p-value WTPS
source           0.9879494  1 0.32024407       0.3229
type             5.8123090  1 0.01591439       0.0204
source:type      3.9517976  1 0.04682133       0.0554
```

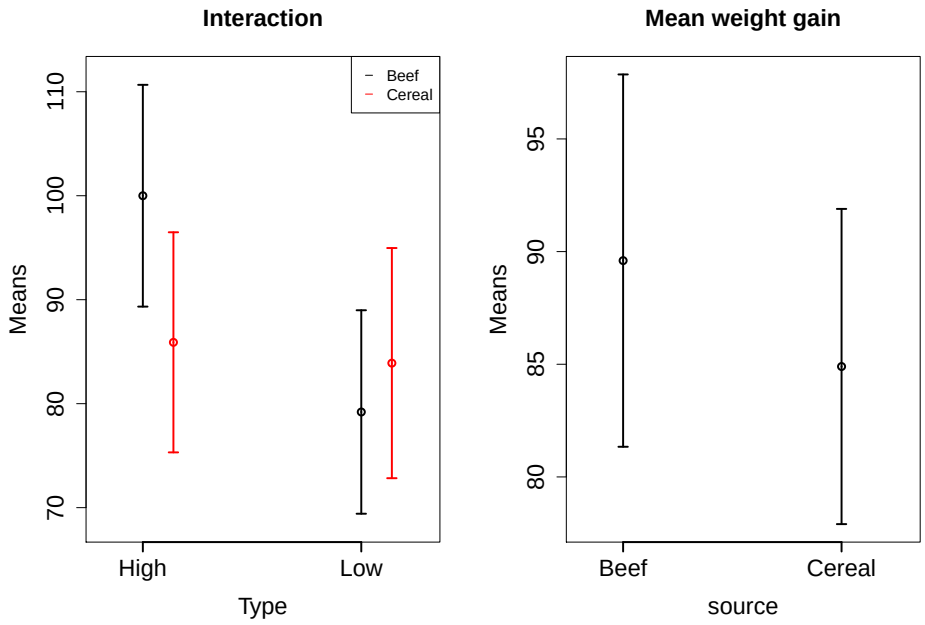

Figure 6: Plots of the interaction of factors `source` and `type` in the weight gain data (left) and for factor `source` alone (right).

```
ANOVA-Type Statistic (ATS):
            Test statistic df1      df2    p-value
source           0.9879494   1 35.72893 0.32692829
type             5.8123090   1 35.72893 0.02118641
source:type      3.9517976   1 35.72893 0.05452616
```

The factor `type`, i.e., high or low amount of protein in the food, has a significant impact on the weight gain at 5% level of significance using all three different tests. The source of the protein, in contrast, does not have a significant influence. The interesting part is the test for interaction: Here, the classical WTS results in a p value of 0.047, whereas both the ATS and WTPS provide a p value of 0.055. Thus, both the ATS and WTPS endorse a "borderline significance" at 5% level.

Figure 6 shows plots for the main effect of the factor `type` as well as the interaction between both factors.

3.5. Three-way layout

For the three-way example, we consider a data set on pizza delivery times (Mackisack 1994). The objective of the study was to see how the delivery time in minutes would be affected by three different factors: whether thick or thin crust was ordered (factor A), whether Coke was ordered with the pizza or not (factor B), and whether or not garlic bread was ordered as a side (factor C). The R code to analyze this data is given in the following statements:

```
R> library("GFD")
R> data("pizza", package = "GFD")
```

```
R> set.seed(1234)
R> model3 <- GFD(Delivery ~ Crust * Coke * Bread, data = pizza)
R> summary(model3)
R> plot(model3, factor = "Crust:Coke:Bread", legendpos = "center",
+    main = "Delivery time of pizza", xlab = "Bread",  cex.axis = 1.5,
+    cex.lab = 1.5, cex.main = 1.5, lwd = 2)
R> plot(model3, factor = "Crust:Coke", legendpos = "topleft",
+    main = "Two-way interaction", xlab = "Coke",  cex.axis = 1.5,
+    cex.lab = 1.5, cex.main = 1.5, lwd = 2)


Call:
Delivery ~ Crust * Coke * Bread

Descriptive:
  Crust Coke Bread n Means Variances Lower 95 % CI Upper 95 % CI
1  thin   no    no 2  19.0       2.0      14.69735      23.30265
5  thin   no   yes 2  17.5       0.5      15.34867      19.65133
3  thin  yes    no 2  17.5       4.5      11.04602      23.95398
7  thin  yes   yes 2  15.0       2.0      10.69735      19.30265
2 thick   no    no 2  19.5       0.5      17.34867      21.65133
6 thick   no   yes 2  18.0       2.0      13.69735      22.30265
4 thick  yes    no 2  21.5       0.5      19.34867      23.65133
8 thick  yes   yes 2  18.5       0.5      16.34867      20.65133

Wald-Type Statistic (WTS):
                Test statistic df      p-value p-value WTPS
Crust                    11.56  1 0.0006738585       0.0089
Coke                      0.36  1 0.5485062355       0.5613
Crust:Coke                6.76  1 0.0093223760       0.0286
Bread                    11.56  1 0.0006738585       0.0073
Crust:Bread               0.04  1 0.8414805811       0.8153
Coke:Bread                1.00  1 0.3173105079       0.3457
Crust:Coke:Bread          0.04  1 0.8414805811       0.8212

ANOVA-Type Statistic (ATS):
                Test statistic df1      df2    p-value
Crust                    11.56   1 4.699248 0.02121110
Coke                      0.36   1 4.699248 0.57625702
Crust:Coke                6.76   1 4.699248 0.05122842
Bread                    11.56   1 4.699248 0.02121110
Crust:Bread               0.04   1 4.699248 0.84984482
Coke:Bread                1.00   1 4.699248 0.36598284
Crust:Coke:Bread          0.04   1 4.699248 0.84984482
```

We find a significant influence of the factors `Crust` and `Bread`. The WTS and WTPS also suggest a significant interaction between the factors `Crust` and `Coke` at 5% level, which is only

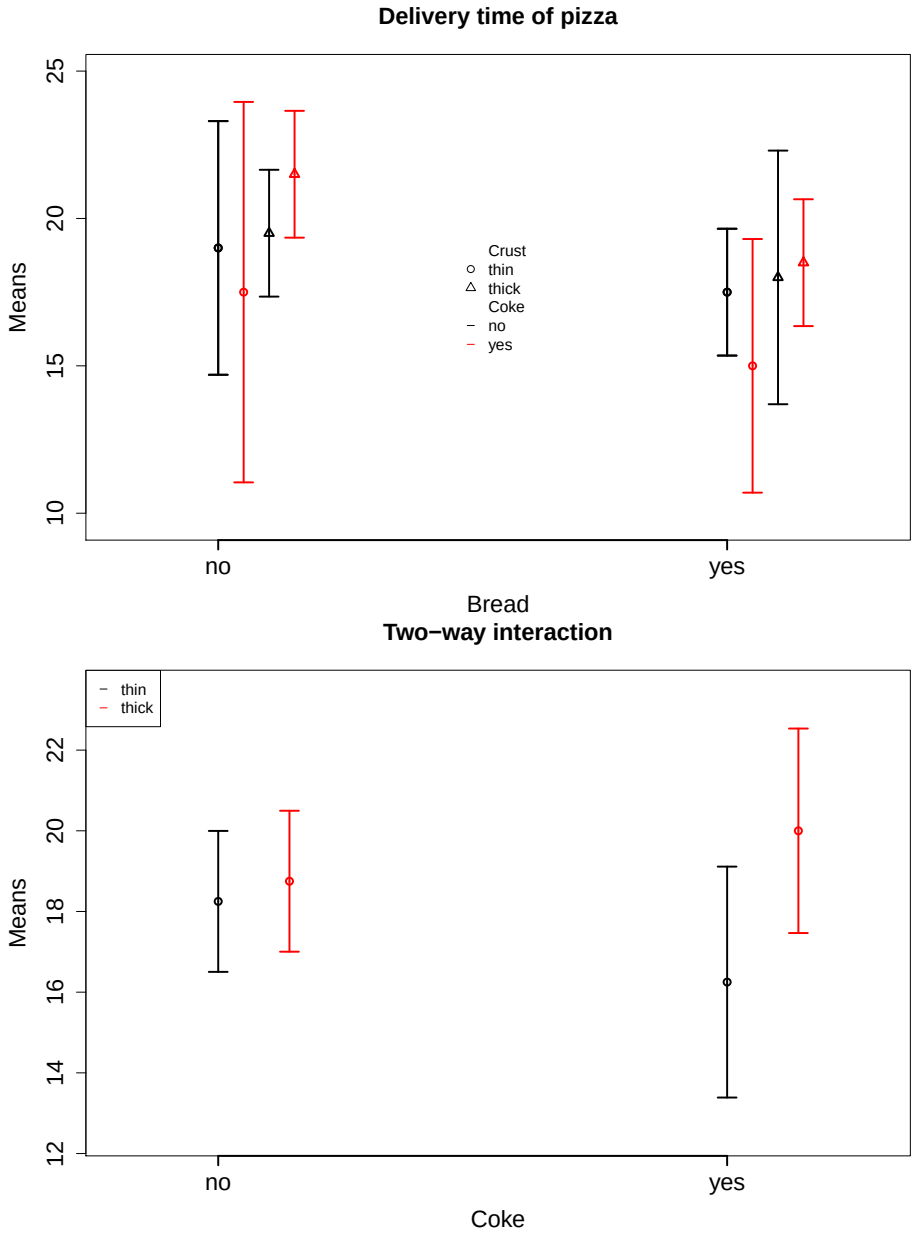

Figure 7: Plots of the three-way interaction (upper panel) and the two-way interaction between factors `Coke` and `Crust` (lower panel).

borderline significant when using the ATS. Figure 7 shows interaction plots of the three-way interaction as well as the two-way interaction between `Crust` and `Coke`.

3.6. Nested design

A nested design is covered by the model

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},$$

where factor B is nested within the levels of factor A. As an example, we consider the `curdies`

data set (Quinn, Lake, and Schreiber 1996) included in the **GFD** package. The aim of the study was to describe basic patterns of variation in a small flatworm, Dugesia, in the Curdies River, Western Victoria. Therefore, worms were sampled at two different seasons and three different sites within each season. For our analyses we consider both factors as fixed (e.g., some sites may only be accessed in summer). The R code for analyzing this nested design is given in the following:

```
R> library("GFD")
R> data("curdies", package = "GFD")
R> set.seed(987)
R> nested <- GFD(dugesia ~ season + season:site, data = curdies)
R> summary(nested)
R> plot(nested, factor="season:site", xlab = "site", cex.axis = 1.5,
+    cex.lab = 1.5, cex.main = 1.5, lwd = 2)

Call:
dugesia ~ season + season:site

Descriptive:
  season  site n     Means  Variances Lower 95 % CI Upper 95 % CI
1 SUMMER     4 6 0.4190947  0.4615290   -0.25954958     1.0977390
2 SUMMER     5 6 0.2290862  0.3148830   -0.33146759     0.7896401
3 SUMMER     6 6 0.1942443  0.0729142   -0.07549781     0.4639864
4 WINTER     1 6 2.0494375  4.0647606    0.03543415     4.0634408
5 WINTER     2 6 4.1819078 35.6801853   -1.78509515    10.1489107
6 WINTER     3 6 0.6782063  0.1910970    0.24151987     1.1148927

Wald-Type Statistic (WTS):
            Test statistic df    p-value p-value WTPS
season            5.415180  1 0.01996239       0.0001
season:site       5.200991  4 0.26728919       0.3154

ANOVA-Type Statistic (ATS):
            Test statistic      df1      df2    p-value
season            5.415180 1.000000 6.447707 0.05593278
season:site       1.382224 1.217424 6.447707 0.29278958
```

In this setting, both WTS and WTPS detect a significant influence of the season whereas the ATS, again, only shows a borderline significance at 5% level. The effect of the site is not significant. A plot for the nested effect is given in Figure 8.

4. Conclusion and future work

The R package **GFD** implements a broad range of semi-parametric methods for the analysis of general factorial designs, i.e., linear models without the assumption of normality and/or homoscedastic variances across the treatment groups. Three different methods are implemented:

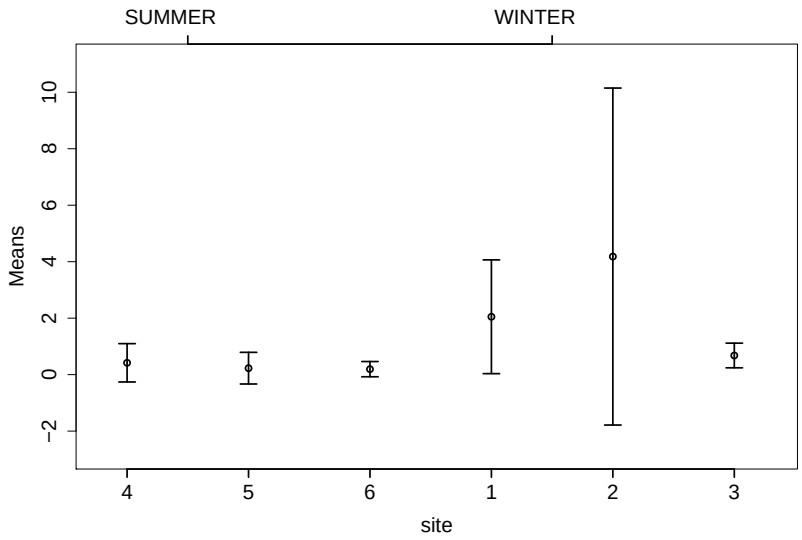

Figure 8: Plot for the effects in the nested design. The sites are nested within seasons.

Wald-type statistic Q_N, ANOVA-type statistic A_N as well as a permutation approach proposed by Pauly *et al.* (2015). All methods can be used to test general hypotheses among the main and interaction effects. In particular, nested designs can be analyzed using **GFD**. From a practical point of view we recommend the WTPS procedure since it has been found in Pauly *et al.* (2015) to posses both good finite type-I error rate control and power behavior. The ATS and WTS, in comparison, are slightly conservative or rather liberal, respectively. Confidence interval plots are available for all effects of interest – except of four- and higher-way interactions.

A graphical user interface (GUI) has been implemented which allows a convenient use of the software in industry, academia, and educational purposes. We plan to update the **GFD** package on a regular basis with new procedures available for the analysis of general designs. So far, ANOVA-based methods are implemented, and an adjustment of the treatment effects for covariates is not possible. Furthermore, tests and simultaneous confidence intervals for multiple comparisons based on the permutation approach are not yet available. The extension of the implemented methods to covariates and multiple comparisons and their implementation will be part of future research.

Acknowledgments

The work of Sarah Friedrich and Markus Pauly was supported by the German Research Foundation project DFG-PA 2409/3-1.

References

Aho K (2017). ***asbio**: A Collection of Statistical Tools for Biologists*. R package version 1.4-2, URL https://CRAN.R-project.org/package=asbio.

Akritas MG, Arnold SF, Brunner E (1997). "Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs." *Journal of the American Statistical Association*, **92**(437), 258–265. doi:10.2307/2291470.

Akritas MG, Brunner E (1997). "A Unified Approach to Rank Tests for Mixed Models." *Journal of Statistical Planning and Inference*, **61**(2), 249–277. doi:10.1016/s0378-3758(96)00177-2.

Bates D, Mächler M (2017). ***Matrix***: *Sparse and Dense Matrix Classes and Methods*. R package version 1.2-10, URL https://CRAN.R-project.org/package=Matrix.

Bathke AC, Schabenberger O, Tobias RD, Madden LV (2009). "Greenhouse-Geisser Adjustment and the ANOVA-Type Statistic: Cousins or Twins?" *The American Statistician*, **63**(3), 239–246. doi:10.1198/tast.2009.08187.

Brunner E, Dette H, Munk A (1997). "Box-Type Approximations in Nonparametric Factorial Designs." *Journal of the American Statistical Association*, **92**(440), 1494–1502. doi:10.1080/01621459.1997.10473671.

Brunner E, Puri ML (2001). "Nonparametric Methods in Factorial Designs." *Statistical Papers*, **42**(1), 1–52. doi:10.1007/s003620000039.

Cengage College (2008). http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/owan/frames/frame.html. [Accessed 28-04-2016].

Chung E, Romano JP (2013). "Exact and Asymptotically Robust Permutation Tests." *The Annals of Statistics*, **41**(2), 484–507. doi:10.1214/13-aos1090.

Everitt BS, Hothorn T (2017). ***HSAUR***: *A Handbook of Statistical Analyses Using R (1st Edition)*. R package version 1.3-8, URL https://CRAN.R-project.org/package=HSAUR.

Fox J, Weisberg S (2011). *An R Companion to Applied Regression*. 2nd edition. Sage, Thousand Oaks.

Hand DJ, Daly F, McConway K, Lunn D, Ostrowski E (1993). *A Handbook of Small Data Sets*. CRC Press.

Hankin RKS (2005). "Recreational Mathematics with R: Introducing The **magic** Package." *R News*, **5**(1), 48–51.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a Class of Permutation Tests: The **coin** Package." *Journal of Statistical Software*, **28**(8), 1–23. doi:10.18637/jss.v028.i08.

Janssen A (1997). "Studentized Permutation Tests for Non-Iid Hypotheses and the Generalized Behrens-Fisher Problem." *Statistics & Probability Letters*, **36**(1), 9–21. doi:10.1016/s0167-7152(97)00043-6.

Janssen A (2005). "Resampling Student's t-Type Statistics." *The Annals of the Institute of Statistical Mathematics*, **57**(3), 507–529. doi:10.1007/bf02509237.

Johansen S (1980). "The Welch-James Approximation to the Distribution of the Residual Sum of Squares in a Weighted Linear Regression." *Biometrika*, **67**(1), 85–92. doi:10.2307/2335320.

Lawrence M, Temple Lang D (2010). "**RGtk2**: A Graphical User Interface Toolkit for R." *Journal of Statistical Software*, **37**(8), 1–52. doi:10.18637/jss.v037.i08.

Lemon J (2006). "**plotrix**: A Package in the Red Light District of R." *R News*, **6**(4), 8–12.

Mackisack M (1994). "What Is the Use of Experiments Conducted by Statistics Students?" *Journal of Statistics Education*, **2**(1), 1–15. URL https://ww2.amstat.org/publications/jse/v2n1/mackisack.html.

Pauly M, Brunner E, Konietschke F (2015). "Asymptotic Permutation Tests in General Factorial Designs." *Journal of the Royal Statistical Society B*, **77**(2), 461–473. doi:10.1111/rssb.12073.

Placzek M, Konietschke F, Pauly M (2014). "Studentisierte Permutationstests für verbundene und vnverbundene 2-Stichprobenprobleme." In *KSFE 2014 – Konferenz der SAS-Anwender in Forschung und Entwicklung*.

Quinn GP, Lake PS, Schreiber ESG (1996). "Littoral Benthos of a Victorian Lake and Its Outlet Stream: Spatial and Temporal Variation." *Australian Journal of Ecology*, **21**(3), 292–301. http://users.monash.edu.au/~murray/AIMS-R-users/downloads/data/curdies.csv, [Accessed 28-04-2016].

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien, Austria. URL https://www.R-project.org/.

Vallejo G, Fernández MP, Livacic-Rojas PE (2010). "Analysis of Unbalanced Factorial Designs with Heteroscedastic Data." *Journal of Statistical Computation and Simulation*, **80**(1), 75–88. doi:10.1080/00949650802482386.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. doi:10.1007/978-0-387-21706-2.

Welch BL (1951). "On the Comparison of Several Mean Values: An Alternative Approach." *Biometrika*, **38**(3–4), 330–336.

Wickham H (2011). "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software*, **40**(1), 1–29. doi:10.18637/jss.v040.i01.

Zhang JT (2012). "An Approximate Degrees of Freedom Test for Heteroscedastic Two-Way ANOVA." *Journal of Statistical Planning and Inference*, **142**(1), 336–346. doi:10.1016/j.jspi.2011.07.023.

Affiliation:

Sarah Friedrich, Markus Pauly
Institute of Statistics
Ulm University
89081 Ulm, Germany
E-mail: sarah.friedrich@uni-ulm.de, markus.pauly@uni-ulm.de

Frank Konietschke
Department of Statistics
University of Texas at Dallas
Dallas, TX 75080, United States of America
Email: fxk141230@utdallas.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
July 2017, Volume 79, Code Snippet 1 *Submitted:* 2015-11-18
doi:10.18637/jss.v079.c01 *Accepted:* 2016-07-14