Journal of Statistical Software

March 2015, Volume 64, Issue 9. http://www.jstatsoft.org/

nparcomp: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals

Frank Konietschke
The University of Texas at Dallas

Marius Placzek
University of Göttingen

Frank Schaarschmidt
Leibniz University of Hannover

Ludwig A. Hothorn
Leibniz University of Hannover

Abstract

One-way layouts, i.e., a single factor with several levels and multiple observations at each level, frequently arise in various fields. Usually not only a global hypothesis is of interest but also multiple comparisons between the different treatment levels. In most practical situations, the distribution of observed data is unknown and there may exist a number of atypical measurements and outliers. Hence, use of parametric and semiparametric procedures that impose restrictive distributional assumptions on observed samples becomes questionable. This, in turn, emphasizes the demand on statistical procedures that enable us to accurately and reliably analyze one-way layouts with minimal conditions on available data. Nonparametric methods offer such a possibility and thus become of particular practical importance. In this article, we introduce a new R package **nparcomp** which provides an easy and user-friendly access to rank-based methods for the analysis of unbalanced one-way layouts. It provides procedures performing multiple comparisons and computing simultaneous confidence intervals for the estimated effects which can be easily visualized. The special case of two samples, the nonparametric Behrens-Fisher problem, is included. We illustrate the implemented procedures by examples from biology and medicine.

Keywords: nonparametric, one-way layout, **nparcomp**, R.

1. Introduction

In many experiments more than two treatment groups are involved. Hereby, the global null hypothesis, i.e., no impact of the treatment on the response, is often not the main question.

Multiple comparisons, e.g., multiple Dunnett-type many-to-one (Dunnett 1955) comparisons, or Tukey-type all-pairs comparisons (Tukey 1953), with an accompanying computation of simultaneous confidence intervals (SCI), are particularly of practical importance. By controlling the familywise error rate in the strong sense, the SCI and the multiple comparison decision, e.g., by multiplicity adjusted p-values, must be compatible with each other. This means, it cannot occur that an individual null hypothesis has been rejected by the multiple comparison procedure, but the corresponding SCI contains the value null coming from the null hypothesis (Konietschke, Hothorn, and Brunner 2012a). It is well known that the classical Bonferroni adjustment can be used to perform multiple comparisons as well as for the computation of compatible SCI. This approach, however, has a low power, particulary when the test statistics are not independent. Bretz, Genz, and Hothorn (2001) propose exact multiple contrast tests (MCTP) and SCI for means of independent and homoscedastic normal samples. The procedures allow for testing arbitrary contrasts, e.g., Dunnett-type, Tukey-type or changepoint comparisons (Hirotsu 1997) and take the correlation between the test statistics into account. Thus, MCTP provide an extensive tool for the computation of compatible SCI. Hereby, the SCI are computed as in the univariate case as "*Mean* $\pm$ $t_{1-\alpha,\nu}(\mathbf{R})$-quantile $\cdot$ *Standard Error*", just by replacing the common univariate $t_{1-\alpha,\nu}$-quantile by an equicoordinate quantile $t_{1-\alpha,\nu}(\mathbf{R})$ coming from a multivariate t distribution with correlation matrix $\mathbf{R}$ and ν degrees of freedom. Multiplicity adjusted p-values for the individual hypotheses are computed by using the cumulative multivariate t distribution function (Genz and Bretz 2009) instead of the univariate t distribution function. Therefore, the results of these procedures can be easily interpreted and are particularly of practical importance. Konietschke, Bösiger, Brunner, and Hothorn (2013) compare exact multiple contrast tests with the ANOVA and conclude that both procedures have comparable power. For a comprehensive overview of parametric multiple contrast tests and SCI we refer the reader to Bretz, Hothorn, and Westfall (2010) and references therein. In particular, parametric methods are numerically available using the R-package **multcomp** (Hothorn, Bretz, and Westfall 2008). We note that the parametric procedures use the critical values from the extreme tail portion of the multivariate t distribution, which is the portion most sensitive to nonnormality. Therefore the problem of robustness will be more serious for SCI compared to individual intervals.

Nonparametric inferences, i.e., without assuming a specific distribution of the data, however, arise in a variety of problems in biomedical research, e.g., in case of skewed data or ordered categorical data. While parametric inferences usually deal with differences between population means, there is an increasing focus in medicine on effect size measures on an individual basis (Browne 2010). For two independent samples, say group 1 and group 2, the relative effect size measure

$$p = P(X < Y) + 1/2P(X = Y) \tag{1}$$

represents the probability that a randomly chosen subject in treatment group 1 reveals a smaller response value X than a randomly chosen subject from treatment group 2 with response value Y. If $p < 1/2$, then the values in group 1 tend to be larger than those in group 2. If $p = 1/2$, none of the observations tend to be smaller or larger.

It is the aim of the present paper to introduce an R extension package called **nparcomp** (Konietschke 2015), which can be used to compute nonparametric MCTP and SCI for relative treatment effects given in (1). We hereby propose algorithms for single step MCTP proposed by Steel (1960), Konietschke *et al.* (2012a) as well as stepwise procedures derived by Gao,

Alvo, Chen, and Li (2008). The package is online available and can be downloaded from the Comprehensive R Archive Network (CRAN), see Konietschke (2015).

The paper is organized as follows. In Section 2 the statistical model, purely nonparametric effects and hypotheses are introduced. In Section 3 the use of multiple contrast test procedures and simultaneous confidence intervals are explained, while Section 4 demonstrates the application of stepwise procedures. The different routines of the software package **nparcomp** are explained in Section 5. The paper closes with a discussion and an outlook for future projects.

2. Statistical model, effects, and hypotheses

We consider a completely randomized one-way layout with a treatment groups and n_i independent replications within the ith treatment group. Without specifying an explicit distribution (e.g., normal distribution) the statistical model can be described by

$$X_{ik} \sim F_i,\ i = 1, \ldots, a;\ k = 1, \ldots, n_i, \tag{2}$$

where $F_i(x) = P(X_{ik} < x) + 1/2P(X_{ik} = x)$ denotes the average of the left and right continuous version of the distribution function. The statistical model does not include any parameters, such as means, which can be used to describe treatment effects. Therefore, the marginal distribution functions are used to describe treatment effects by

$$p_i = \int H dF_i = P(Z < X_{i1}) + 1/2P(Z = X_{i1}),\ i = 1, \ldots, a, \tag{3}$$

where $H = \frac{1}{a}\sum_{j=1}^{a} F_j$ denotes a mean distribution in its unweighted (Brunner and Puri 2001; Gao *et al.* 2008) form. Here Z represents a random variable with distribution H being independently distributed from X_{i1}. These effects are called *unweighted relative effects* (Gao *et al.* 2008; Konietschke *et al.* 2012a). They can be interpreted as the probability that an observation Z - randomly chosen from all observations - has a smaller value than a randomly chosen observation from sample i. In the case of $p_i > 1/2$ data from sample i tend to larger values than Z. If $p_i = 1/2$, neither X_{i1} nor Z tends to larger or smaller values. In particular, if $p_i < p_j$, then the values in group i tend to be smaller than those in group j; if $p_i = p_j$, none of the observations tend to be smaller or larger. Figure 1 illustrates these relations for two normal distributions.

In the special case of independent ordinal data p_i is also called *ordinal effect size measure* (Ryu and Agresti 2008; Ryu 2009).

2.1. Hypotheses

Gao *et al.* (2008) propose rank based multiple stepwise procedures for testing the Dunnett-type (Dunnett 1955) multiple comparisons

$$H_0^F : \begin{cases} F_1 = F_2 \\ F_1 = F_3 \\ \vdots \\ F_1 = F_a \end{cases} \iff H_0^F : \mathbf{CF} = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ -1 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & \ldots & \ldots & 1 \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_a \end{pmatrix} = \mathbf{0}, \tag{4}$$

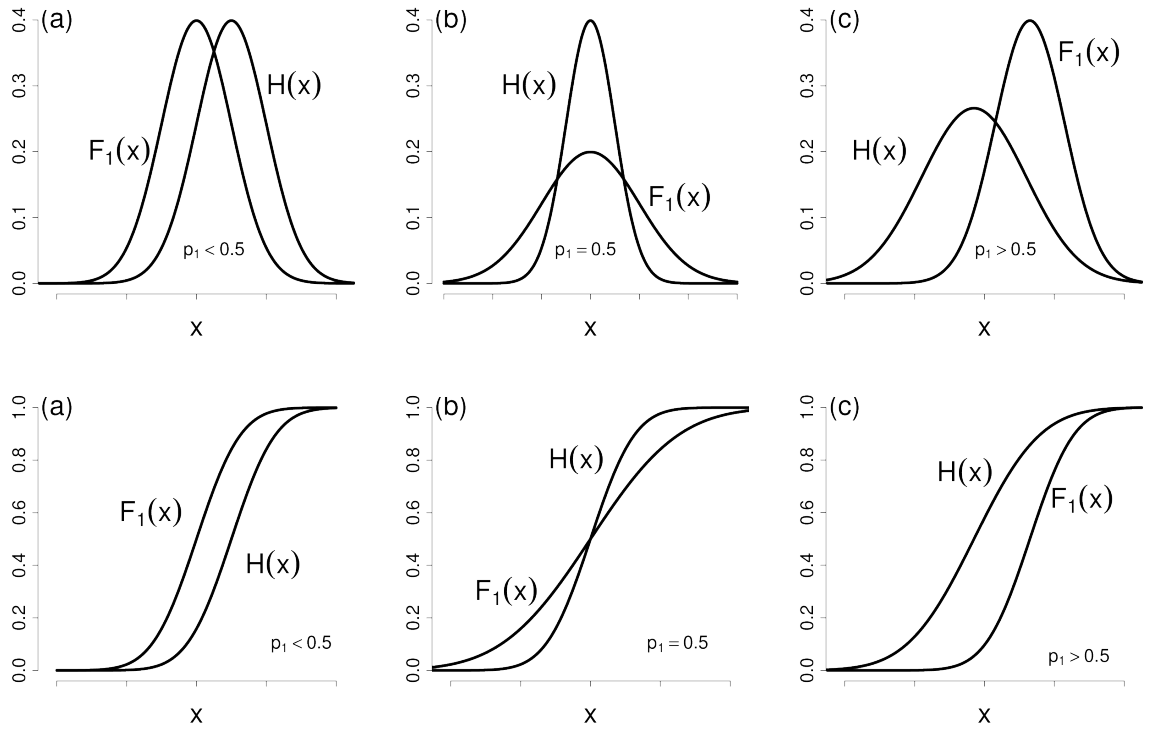

Figure 1: Interpretation of the nonparametric relative effects: density functions (top), distribution functions (bottom).

as well as the Tukey-type (Tukey 1953) multiple comparisons

$$H_0^F : \begin{cases} F_1 = F_2 \\ F_1 = F_3 \\ \quad \vdots \\ F_1 = F_a \\ F_2 = F_3 \\ \quad \vdots \\ F_{a-1} = F_a \end{cases} \iff H_0^F : \mathbf{CF} = \begin{pmatrix} -1 & 1 & 0 & \dots & \dots & 0 & 0 \\ -1 & 0 & 1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & 0 & \dots & \dots & 1 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & -1 & 1 \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ \vdots \\ F_a \end{pmatrix} = \mathbf{0}, \quad (5)$$

formulated in terms of the distribution functions $F_1, \dots, F_a$ of the data. All test procedures for H_0^F, however, are limited to testing problems and cannot be used to construct confidence intervals for the underlying treatment effects. Therefore, Konietschke *et al.* (2012a) propose multiple contrast test procedures and SCI for the effects $\mathbf{p}$. The procedures allow for an arbitrary user-defined contrast matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^\top \\ \vdots \\ \mathbf{c}_q^\top \end{pmatrix} = \begin{pmatrix} c_{11} & \dots & c_{1a} \\ \vdots & \dots & \vdots \\ c_{1q} & \dots & c_{qa} \end{pmatrix}, \quad (6)$$

where each row vector $\mathbf{c}_\ell^\top$ of $\mathbf{C}$ is one contrast, i.e., each row sum of the contrast matrix is

zero by definition. For example, multiple comparisons to a control are expressed by

$$H_0^p : \begin{cases} p_1 = p_2 \\ p_1 = p_3 \\ \vdots \\ p_1 = p_a \end{cases} \iff H_0^p : \mathbf{Cp} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & \dots & \dots & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_a \end{pmatrix} = \mathbf{0}, \tag{7}$$

all-pairwise comparisons are formulated by

$$H_0^p : \begin{cases} p_1 = p_2 \\ p_1 = p_3 \\ \vdots \\ p_1 = p_a \\ p_2 = p_3 \\ \vdots \\ p_{a-1} = p_a \end{cases} \iff H_0^p : \mathbf{Cp} = \begin{pmatrix} -1 & 1 & 0 & \dots & \dots & 0 & 0 \\ -1 & 0 & 1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & 0 & \dots & \dots & 1 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & \dots & \dots & -1 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ \vdots \\ p_a \end{pmatrix} = \mathbf{0}, \tag{8}$$

and Williams-type (Williams 1972; Bretz 2006; Konietschke and Hothorn 2012) comparisons are expressed by using the contrast matrix

$$H_0^p : \mathbf{Cp} = \begin{pmatrix} -1 & 0 & 0 & \dots & 0 & 1 \\ -1 & 0 & 0 & \dots & \frac{n_{a-1}}{n_{a-1}+n_a} & \frac{n_a}{n_{a-1}+n_a} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & \frac{n_2}{n_2+\dots+n_a} & 0 & \dots & \dots & \frac{n_a}{n_2+\dots+n_a} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_a \end{pmatrix} = \mathbf{0}. \tag{9}$$

For a comprehensive overview of different kinds of contrasts we refer the reader to Bretz *et al.* (2001). We note that the hypothesis in the classical Behrens-Fisher model is contained in this general setup as a special case. This is easily seen from the fact that $p_i = 1/2$ if H and F_i are both symmetric distributions with the same center of symmetry. The nonparametric hypothesis $H_0^F : \mathbf{CF} = \mathbf{0}$ is very general and implies $H_0^p : \mathbf{Cp} = \mathbf{0}$: $H_0^F : \mathbf{CF} = \mathbf{0} \Rightarrow H_0^p : \mathbf{Cp} = \mathbf{C} \int H d\mathbf{F} = \int H d\mathbf{CF} = \mathbf{0}$. The shape of the distribution functions can differ even under the null hypothesis. In the special case of quite restrictive location models $F_i(x) = F(x - \mu_i)$, $i = 1, \dots, a$, the nonparametric and parametric hypotheses in terms of the location parameters μ_i are equivalent. For a detailed discussion of the hypotheses formulated above we refer to Akritas, Arnold, and Brunner (1997) and Brunner and Munzel (2000). Furthermore, a nonparametric procedure for testing independence in distribution for categorical variables between two or more populations is suggested in Finos and Salmaso (2004).

3. Multiple contrast test procedures for H_0^p and SCI

Gao *et al.* (2008) and Konietschke *et al.* (2012a) propose rank-based estimators $\widehat{\mathbf{p}} = (\widehat{p}_1, \dots, \widehat{p}_a)^\top$ for the unweighted treatment effects $\mathbf{p} = (p_1, \dots, p_a)^\top$. In particular, Konietschke *et al.* (2012a) derive the asymptotic distribution of $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$ for arbitrary $\mathbf{p}$. In order to test the individual null hypothesis $H_0^{(\ell)} : \mathbf{c}_\ell^\top \mathbf{p} = 0$, consider the pivotal test statistic

$T_\ell^p = \sqrt{N}\frac{\mathbf{c}_\ell^\top(\widehat{\mathbf{p}}-\mathbf{p})}{\widehat{\sigma}_\ell}$, where $\widehat{\sigma}_\ell$ denotes a consistent estimator of $Var(\sqrt{N}\mathbf{c}_\ell^\top(\widehat{\mathbf{p}}-\mathbf{p}))$. The test statistics T_ℓ^p are collected in the vector

$$\mathbf{T} = (T_1^p, \ldots, T_q^p)^\top. \tag{10}$$

It can be shown that $\mathbf{T}$ follows a multivariate normal distribution with expectation $\mathbf{0}$ and correlation matrix $\mathbf{R} = (r_{\ell m})_{\ell,m=1,\ldots,q}$, asymptotically. Since $\mathbf{R}$ is unknown, it must be replaced by a consistent estimator $\widehat{\mathbf{R}}$ (Konietschke *et al.* 2012a). The individual null hypothesis $H_0^{(\ell)} : \mathbf{c}_\ell^\top\mathbf{p} = 0$ will be rejected at a two-sided multiple level α, if $|T_\ell^{0.5}| \geq t_{1-\alpha,\nu}(\widehat{\mathbf{R}})$. Approximate $(1-\alpha)$-simultaneous confidence intervals for the treatment effects $\delta_\ell = \mathbf{c}_\ell^\top\mathbf{p}$ are obtained from

$$\left[\mathbf{c}_\ell^\top\widehat{\mathbf{p}} - t_{1-\alpha,\nu}(\widehat{\mathbf{R}})\widehat{\sigma}_\ell/\sqrt{N}\ ;\ \mathbf{c}_\ell^\top\widehat{\mathbf{p}} + t_{1-\alpha,\nu}(\widehat{\mathbf{R}})\widehat{\sigma}_\ell/\sqrt{N}\right], \tag{11}$$

where $t_{1-\alpha,\nu}(\widehat{\mathbf{R}})$ denotes the two-sided equicoordinate quantile from the multivariate t - distribution with approximated degree of freedom ν and correlation matrix $\widehat{\mathbf{R}}$ (Konietschke *et al.* 2012a). By construction, the test decision for $H_0^{(\ell)} : \mathbf{c}_\ell^\top\mathbf{p} = 0$ and the SCI are compatible, i.e., it cannot occur that an individual null hypothesis $H_0^{(\ell)} : \mathbf{c}_\ell^\top\mathbf{p} = 0$ is rejected, but the corresponding SCI contains zero. The global null hypothesis $H_0^p : \mathbf{Cp} = \mathbf{0}$ will be rejected, if

$$T_0 = \{|T_1^{0.5}|, \ldots, |T_q^{0.5}|\} \geq t_{1-\alpha,\nu}(\widehat{\mathbf{R}}). \tag{12}$$

One-sided confidence intervals can be computed by replacing the two-sided quantile $t_{1-\alpha,\nu}$ with its one-sided version (Konietschke and Hothorn 2012; Konietschke *et al.* 2012a). Note that the SCI defined in (11) may be not range preserving, i.e., the lower bounds can be smaller than -1 and the upper bounds can be larger than 1. Konietschke *et al.* (2012a) therefore propose the Fisher-approximation for the construction of range-preserving SCI. We note that the rank-based MCTP's control the familywise error rate in the strong sense, which follows from the fact that the set of hypotheses $\mathbf{C}$ and the corresponding test statistics $\mathbf{T}$ constitute a joint-testing family. The proofs are given in Konietschke *et al.* (2012a).

3.1. The two sample problem

In particular, inference methods for testing the null hypothesis $H_0 : p_1 = p_2$ with two-independent samples $X_{ik} \sim F_i,\ i = 1, 2;\ k = 1, \ldots, n_i$, occur frequently in practical applications. Testing the null hypothesis $H_0^F : F_1 = F_2$ can be realized with the Wilcoxon-Mann Whitney test. We note that the relative treatment effect

$$p = p_1 - p_2 + 1/2 = P(X_{11} < X_{21}) + 1/2P(X_{11} = X_{21}). \tag{13}$$

can be easily rewritten as given in (1). If $p > 1/2$, the observations in sample 2 tend to be larger than the observations in sample 1. No data tend to be larger or smaller if $p = 1/2$. Therefore the hypothesis of no treatment effect is formulated by $H_0^p : p = 1/2$, which is known as the nonparametric Behrens-Fisher problem (Brunner and Munzel 2000). We note that the testing problem $H_0^p : p = 1/2$ is not equivalent to location-scale testing problem $H_0 : F_1(t) = F_2(t)$ versus $H_1 : F_2(t) = F_1\left(\frac{t-\mu}{\sigma}\right)$ with $\mu \neq 0$ or $\sigma \neq 1$. Location-scale testing implies that different variances and /or locations may result in significant treatment effects. In the Behrens-Fisher situation one is interested in testing location effects only. Here, even

under the null hypothesis, the marginal distribution functions in the different groups may have different shapes, and are not assumed to be equal. Inference methods for location-scale problems are considered by Marozzi (2009, 2013).

The package **nparcomp** provides the function `npar.t.test` which can be used to test the null hypothesis $H_0^p : p = 1/2$ by using the approximate Brunner-Munzel test (Brunner and Munzel 2000) or by using the approximate studentized permutation test proposed by Neubert and Brunner (2007). Further $(1 - \alpha)$-range preserving confidence intervals using the logit or probit transformation are available (Konietschke 2009). As only asymptotic and approximate procedures are provided by the package, we recommend the use of these procedures with medium sample sizes $n_i \geq 8$.

4. Stepwise procedures for H_0^F

Gao *et al.* (2008) have shown that the vector of estimates $\sqrt{N}\mathbf{C}\widehat{\mathbf{p}}$ is asymptotically multivariate normal with mean $\mathbf{0}$ and a certain covariance matrix $\mathbf{\Sigma}$ under the null hypothesis H_0^F. To address the problem of multiple comparisons to a control as described in (4), Gao *et al.* (2008) have shown that the distribution of the corresponding vector of test statistics $T_\ell^F = \sqrt{N}\mathbf{c}_\ell^\top \widehat{\mathbf{p}}/\widehat{\sigma}_\ell, \ell = 1, \ldots, a-1$, satisfies the multivariate totally positive of order two (MTP2) condition. Therefore, Hochberg's step-up procedure (Hochberg 1986) is applicable to correct the individual p-values for multiplicity. This procedure rejects the individual null hypothesis $H_0^{(\ell^\top)} : \mathbf{c}_{\ell^\top}^\top \mathbf{F} = 0$ $(\ell^\top \leq \ell)$ at significance level α, if $P_{(\ell)} \leq \frac{\alpha}{a-\ell}$.

With regard to all pairwise comparisons defined in (5), Gao *et al.* (2008) generalize various single step and stagewise procedures for H_0^F. Hereby, the modified Campbell and Skillings (Campbell and Skillings 1985) procedure is recommended. At the initial step the treatments are ordered and labelled $1, \ldots, a$ according to their effects $\widehat{p}_i$ and this same labelling is used in the subsequent steps. At the $(a - p + 1)$th step $(p = 2, \ldots, a)$, subsets of the form $\{j, j + 1, \ldots, j + p - 1\}$ are tested if and only if they have not been retained as homogeneous by implication at a previous step. For further details we refer to Gao *et al.* (2008).

5. Software

In this section we present the package **nparcomp** and provide examples that illustrate how the contained functions can be used to analyze the introduced two-sample problem or perform multiple comparisons via single step or stepwise procedures.

5.1. Nonparametric Behrens-Fisher problem

In this setting we analyze a two-sample design where we neither assume homogeneous variances nor any similarities in the shape of the distribution functions. By means of a data example, the functionality will be described explicitely. Consider the *numbers of implantations* data set available in the **nparcomp** package. In a fertility trial with 29 female Wistar rats the experimenter wanted to test if an active treatment influences the fertiliy of the rats. Therefore $n_1 = 12$ rats received a control while $n_2 = 17$ rats were administered a treatment. A first step in the analysis could be to test whether the numbers of implantations in the treatment group differ from the numbers of implantations in the control group. To do so we can use the function `npar.t.test` in the following way, testing $H_0^p : p = 1/2$ versus

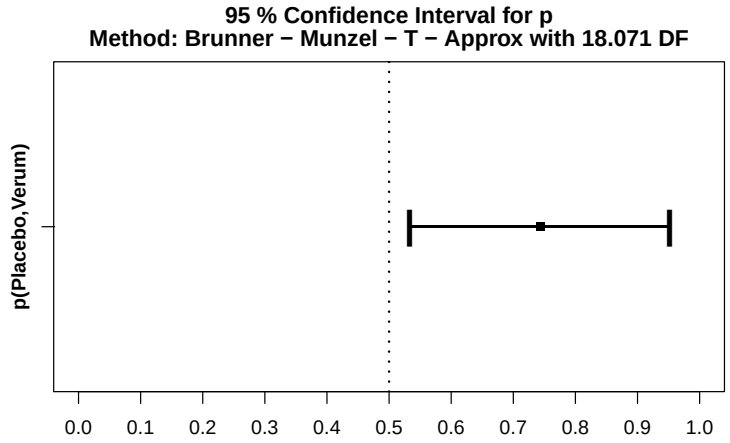

Figure 2: Relative effect $\widehat{p}$ and corresponding 95%-confidence interval in the fertility trial.

$H_1^p : p \neq 1/2$ where p denotes the relative effect between the two treatment groups. By default the `alternative` is set to `"two.sided"`. To obtain one-sided tests one can choose between `"less"` and `"greater"`. We specify `method = "t.app"` to get an approximation by a t distribution as the asymptotic method. Here `npar.t.test` also provides the options `"logit"`, `"probit"` and `"normal"` performing logit/probit transformations or a normal approximation. A studentized permutation test (Neubert and Brunner 2007) can be obtained by setting `method = "permu"`. The confidence level is selected via the parameter `conf.level`. To get a structured overview of the outcome one can use the S3 method `summary`:

```
R> library("nparcomp")
R> data("impla")
R> fert.trial <- npar.t.test(impla ~ group, data = impla,
+    conf.level = 0.95, method = "t.app", info = FALSE)
R> summary(fert.trial)

#---Nonparametric Test Procedures and Confidence Intervals for relative  effects---#

- Alternative Hypothesis:  True relative effect p is less or equal than 1/2
- Confidence level: 95 %
- Method = Brunner - Munzel - T - Approx with 18.071 DF
#---------------------------Interpretation--------------------------------------#
p(a,b) > 1/2 : b tends to be larger than a
#-------------------------------------------------------------------------------#

#----Data Info----------------------------------------------------------------#
        Sample Size
Placebo Placebo   12
Verum     Verum   17

#----Analysis----------------------------------------------------------------#
            Effect Estimator Lower Upper     T p.Value
1 p(Placebo,Verum)     0.743 0.533 0.952 2.429   0.026
```

The numbers of implantations tend to be larger in the verum group ($\widehat{p} = 0.743$). The null hypothesis $H_0 : p = 1/2$ is significantly rejected at 5% level of significance. The **nparcomp**

package provides a S3 method to visualize the estimator and its confidence interval:

```
R> plot(fert.trial)
```

The obtained plot is shown in Figure 2.

5.2. Simultaneous inferences: Single step procedures

Here we use two examples how to analyze a one-way layout with $a \geq 3$ groups to demonstrate the single step procedure described in Section 3. First, consider the *relative liver weights* trial, which was originally analyzed by Brunner and Munzel (2002), p. 97. The data is contained in the package. The response variable is the relative liver weight from $n_1 = 8$ rats in the negative control and $n_2 = 7$, $n_3 = 8$, $n_4 = 7$ and $n_5 = 8$ rats in the dose groups. The interest is in simultaneous many-to-one comparisons, i.e., to test the null hypotheses $H_0^p : p_1 = p_j$, $j = 2, 3, 4, 5$, simultaneously. The function `mctp` provides different kinds of contrast matrices: `"Tukey"` for all-pairs comparisons, `"Dunnett"` for many-to-one comparisons, `"Sequen"` for sequential contrasts, `"Williams"` for Williams-type trend contrasts, `AVE` for average contrasts. In addition `"Changepoint"`, `"McDermott"`, `"Marcus"` and `"UmbrellaWilliams"` contrast are available (Konietschke, Libiger, and Hothorn 2012b). The user can also enter a `"UserDefined"` $q \times a$ contrast matrix containing the contrast coefficients in argument `contrast.matrix`. We apply `mctp` to the data set specifying `type = "Dunnett"` to perform many-to-one comparisons. Just like in the case of `npar.t.test` the `alternative` is set to `"two.sided"` by default. To obtain one-sided tests one has to choose between `"less"` and `"greater"`. There are three options setting the asymptotic method: `"mult.t"` for a multivariate t distribution, `"fisher"` for the Fisher-approximation, `"normal"` for a normal approximation. The confidence level is set via `conf.level`:

```
R> data("liver")
R> tox.trial <- mctp(weight ~ dosage, data = liver, type = "Dunnett",
+    conf.level = 0.95, asy.method = "fisher", info = FALSE)
R> summary(tox.trial)

#-------------Nonparametric Multiple Comparisons for relative effects-------------#

- Alternative Hypothesis:  True differences of relative effects are less or equal
  than 0
- Estimation Method: Global Pseudo ranks
- Type of Contrast : Dunnett
- Confidence Level: 95 %
- Method = Fisher with 11 DF

#-----------------------------------------------------------------------------#

#----Data Info---------------------------------------------------------------#
  Sample Size    Effect     Lower     Upper
1      1    8 0.2738839 0.1883282 0.3801054
2      2    7 0.3168367 0.2033613 0.4572850
3      3    8 0.3618304 0.2750676 0.4586457
4      4    7 0.6938776 0.6123359 0.7648541
5      5    8 0.8535714 0.8002739 0.8945209

#----Contrast---------------------------------------------------------------#
       1 2 3 4 5
```

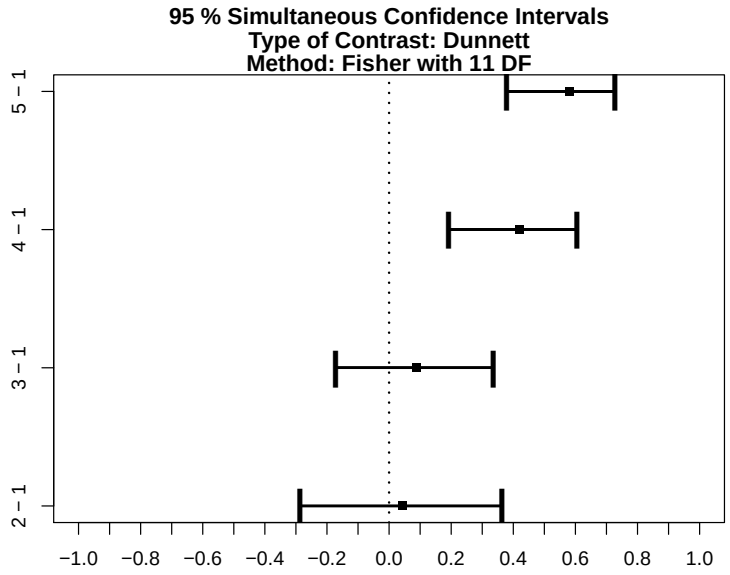

Figure 3: Differences of relative effects and 95%-confidence intervals in the toxicity trial.

```
2 - 1 -1 1 0 0 0
3 - 1 -1 0 1 0 0
4 - 1 -1 0 0 1 0
5 - 1 -1 0 0 0 1

#----Analysis-------------------------------------------------------------------------#
      Estimator  Lower Upper Statistic      p.Value
2 - 1     0.043 -0.288 0.364     0.353 0.9818154758
3 - 1     0.088 -0.172 0.337     0.937 0.7114134844
4 - 1     0.420  0.192 0.605     4.922 0.0011974456
5 - 1     0.580  0.379 0.728     7.003 0.0003226421

#----Overall--------------------------------------------------------------------------#
  Quantile      p.Value
1 2.784112 0.0003226421
#-------------------------------------------------------------------------------------#
```

The relative liver weights tend to larger values for increasing dose ($\widehat{\mathbf{p}} = (0.27, 0.32, 0.36, 0.69, 0.85)^\top$). Simultaneous 95%-confidence intervals for the effects $\delta_j = p_j - p_1$ as well as multiplicity adjusted p-values are displayed in the `Analysis` section of the output. Here, significant differences at 5% level occur between the negative control and both groups 4 and 5, respectively. A confidence interval plot is achieved via:

```
R> plot(tox.trial)
```

For each individual null hypothesis the estimator and its 95% confidence interval is plotted (Figure 3). Here we can see the advantage of simultaneous confidence intervals. Whenever an individual hypothesis has been rejected by the multiple contrast test, the corresponding simultaneous confidence interval does not include the null.

In the second example we will analyze the *colorectal cancer* data set (Ryu 2009) contained in the package. It consists of 174 patients suffering from colorectal cancer which were randomly

assigned to one out of three treatment regimens: IFL (irinotecan, bolus fluorouracil, and leucovorin), FOLFOX (infused fluorouracil, leucovorin, and ocaliplatin), and IROX (irinotecan and oxaliplatin). The response variable is the patients' appetite scores at their third visit for each treatment. There are five ordered categories: (1) normal, (2) not always good, (3) don't really enjoy, (4) force to eat, (5) can't stand. Ryu (2009) already analyzed this data set when presenting a score method with studentized range distribution and accompanied simultaneous confidence intervals for the ordinal effect measures. We will compare our results with the proposed method.

The interest is in finding statistical differences between two treatment regimens. Therefore simultaneous all-pairs comparisons, i.e., testing the null hypotheses $H_0^p : p_i = p_j$, $j = 1, 2, 3$, simultaneously, will be performed. To this end we apply `mctp` to the data set specifying `type = "Tukey"` to stay compatible with Ryu (2009):

```
R> data("appetite")
R> col.cancer <- mctp(Score ~ Group, data = appetite, type = "Tukey",
+    conf.level = 0.95, asy.method = "fisher", info = FALSE)
R> summary(col.cancer)


#--------------Nonparametric Multiple Comparisons for relative effects-------------#

- Alternative Hypothesis:  True differences of relative effects are less or equal
  than 0
- Estimation Method: Global Pseudo ranks
- Type of Contrast : Tukey
- Confidence Level: 95 %
- Method = Fisher with 104 DF

#-----------------------------------------------------------------------------------#

#----Data Info-------------------------------------------------------------------#
       Sample Size    Effect     Lower     Upper
FOLFOX FOLFOX   53 0.5748523 0.5275324 0.6208392
IFL       IFL   66 0.4170825 0.3769846 0.4583080
IROX     IROX   55 0.5080652 0.4630222 0.5529777

#----Contrast-------------------------------------------------------------------#
              FOLFOX IFL IROX
IFL - FOLFOX      -1   1    0
IROX - FOLFOX     -1   0    1
IROX - IFL         0  -1    1

#----Analysis-------------------------------------------------------------------#
              Estimator  Lower  Upper Statistic     p.Value
IFL - FOLFOX     -0.158 -0.264 -0.048    -3.390 0.002794005
IROX - FOLFOX    -0.067 -0.184  0.053    -1.330 0.381027703
IROX - IFL        0.091 -0.014  0.194     2.064 0.101981552

#----Overall--------------------------------------------------------------------#
  Quantile     p.Value
1  2.37573 0.002794005
#-----------------------------------------------------------------------------------#
```

The appetite scores in the IFL group ($\widehat{p}_{\text{IFL}} = 0.417$) tend to be smaller than those in the other two groups ($\widehat{p}_{\text{IROX}} = 0.508$, $\widehat{p}_{\text{FOLFOX}} = 0.574$). The `Analysis` section contains simultaneous 95%-confidence intervals for the effects $\delta_{ij} = p_i - p_j,\ i > j$, as well as multiplicity adjusted p-values. A significant difference at 5% level occurs only between IFL and FOLFOX. This means the FOLFOX regimen leads to higher appetite scores than the IFL treatment. Thus one can infer that FOLFOX causes less severe appetite problems than IFL. We further note the three groups are not genuine random samples taken from a parent distribution, but that they are obtained through randomization of a non random sample. Therefore caution should be paid when drawing inferences (see Pesarin and Salmaso (2010), p. 10ff.).

Ryu (2009) applies a score method based on pairwise relative effects to this dataset and obtains the same decisions. Only the 95%-confidence interval for the comparison between IFL and FOLFOX ([0.546,0.751]) does not include 1/2. However, note that an analysis based on pairwise effects can lead to paradox decisions and therefore global rank based procedures are preferred. We further note that Munzel and Hothorn (2001) propose non-parametric multiple contrast tests and SCI for pairwisely defined relative effects. In particular, Munzel and Hothorn (2001) implements the results in the R package **npmc**, which was, however, removed from CRAN due to abundance.

5.3. Simultaneous inferences: Stepwise procedures

In the previous section we have already seen how to perform many-to-one comparisons applying the single-step procedure `mctp` to the *relative liver weights* data set. Now we want to show an example of how to analyze a one-way layout with $a \geq 3$ groups using the stepwise procedure introduced by Gao *et al.* (2008) presented in Section 4. The function `gao` implements a stepwise procedure which can only be used to perform nonparametric multiple tests for many-to-one comparisons. It is contained in the **nparcomp** package and can be called as follows. By default the first group by lexicographical ordering is handeled as control group. However, the user can specify it via the parameter `control` entering a character string defining the control group:

```
R> data("liver")
R> gao(weight ~ dosage, data = liver, alpha = 0.05)

#----Xin Gao et al's (2008) Non-Parametric Multiple Test Procedure
#----Type of Adjustment: Hochberg
#----Level of significance = 0.05
#----The procedure compares if the distribution functions F() are equal. The FWER
     is strongly controlled
#---- This function uses pseudo ranks of the data!
#----Reference: Gao, X. et al. (2008). Nonparametric Multiple Comparison Procedures
     for Unbalanced One-Way Factorial Designs. JSPI 138, 2574 - 2591.

$Info
  Sample Size Effect Variance
1      1    8 0.2739   0.0406
2      2    7 0.3168   0.0653
3      3    8 0.3618   0.0279
4      4    7 0.6939   0.0273
5      5    8 0.8536   0.0121

$Analysis
```

```
  Comparison Estimator      df Statistic  P.Raw P.Hochberg Rejected P.Bonf  P.Holm
1  F(2)-F(1)    0.0430 11.4072    0.3580 0.7269     0.7269    FALSE 1.0000  0.7269
2  F(3)-F(1)    0.0879 13.5348    0.9508 0.3584     0.7167    FALSE 1.0000  0.7167
3  F(4)-F(1)    0.4200 12.9622    4.4348 0.0007     0.0020     TRUE 0.0027  0.0020
4  F(5)-F(1)    0.5797 10.8470    7.1413 0.0000     0.0001     TRUE 0.0001  0.0001
```

Note that in contrast to `mctp` - which tested via relative effects - `gao` tests the hypothesis of no treatment effect in terms of distribution functions. Thus, the overall hypthesis is H_0^F : $F_1 = F_j \ \forall j = 2, 3, 4, 5$. It has to be rejected because the smallest p-value is less than 0.05. For each of the many-to-one comparison the analysis table contains a raw p-value and adjusted p-values obtained by the Hochberg-adjustment, Bonferroni-adjustment and Holm's procedure. Consider the adjusted p-values of the Hochberg-adjustment. According to these values the first two dose levels have no significant effect on the relative liver weight. However, the hypotheses of no treatment effect of dose levels 3 and 4 are rejected. The multiple level is strongly controlled by the adjustment. Comparing these results with those obtained by applying `mctp` to the data, we can state that in this example both analysis are consonant in their decisions.

nparcomp also provides the function `gao_cs` – the implementation of the Gao *et al.* (2008) modification of the Campbell and Skillings (1985, CS) stepwise multiple comparison procedure for all-pairs comparisons. Its usage shall be explained with the help of the *reaction* data set available in the **nparcomp** package and taken from Shirley (1977). The data set contains the results of a toxicity trial including four dose groups abbreviated with 0 through 3. The response variable is the reaction time in seconds of $N = 40$ mice. It is a balanced design with $n \equiv 10$. The interest is in all-pairs comparisons, i.e., $H_0^F : F_i = F_j$, $i, j = 0, 1, 2, 3$, to check whether the increasing dose levels of the active treatment influence the reaction time of the mice.

```
R> data("reaction")
R> gao_cs(Time ~ Group, data = reaction, alpha = 0.05)

#----Gao et al's (2008) modification of Campbell and Skillings (1985) (CS)
     stepwise multiple comparison procedure
#---- This function uses joint ranks of the data. Attention: In the CS algorithm,
      the samples are jointly reranked!
#----Reference: Gao, X. et al. (2008). Nonparametric Multiple Comparison Procedures
     for Unbalanced One-Way Factorial Designs. JSPI 138, 2574 - 2591.
$Info
  Order Sample Size  Effect   Variance
1     1      0   10 0.19375 0.01643229
2     2      1   10 0.50500 0.05757639
3     3      2   10 0.57875 0.07524479
4     4      3   10 0.72250 0.05256250

$Single.Analysis
  Comp Effect Statistic      DF  P.RAW p.BONF p.HOLM
1  3-0 0.5288    6.3656 14.1262 0.0000 0.0001 0.0001
2  2-0 0.3850    4.0210 12.7520 0.0015 0.0090 0.0075
3  3-1 0.2175    2.0725 17.9628 0.0529 0.3173 0.1587
4  1-0 0.3113    3.6180 13.7503 0.0029 0.0172 0.0115
5  2-1 0.0737    0.6399 17.6870 0.5304 1.0000 0.5304
6  3-2 0.1438    1.2715 17.4504 0.2202 1.0000 0.4404
```

```
$CS.Analysis
  Comp Effect Statistic      DF Quantiles  Adj.P  Alpha Rejected Layer
1  3-0 0.5288    9.0024 14.1262    4.1059  1e-04 0.0500     TRUE     1
2  2-0 0.4200    5.9397 14.1745    3.6962 0.0023 0.0500     TRUE     2
3  3-1 0.2650    3.1379 17.9953    3.6094 0.0949 0.0500    FALSE     2
4  1-0 0.3800    5.3725 16.8977    3.4699 0.0014 0.0253     TRUE     3
5  2-1 0.1000    1.0594 17.8029     3.453 0.4636 0.0253    FALSE     3
6  3-2 0.1750    1.9174 17.7942     3.453 0.1921 0.0253    FALSE     3
```

The output mainly consists of two sections – the *single analysis* which contains raw p-values accompanied by Bonferroni- and Holm-adjusted p-values and the *CS analysis* which realizes the adjustment obtained by modifying Campbell and Skillings stepwise procedure. In simulations the CS-adjustment shows the best performence and is therefore recommended for real data evaluations. Here, the overall null hypothesis, i.e., $H_0^F : F_i = F_j$, $\forall i, j = 0, 1, 2, 3$, is rejected, because three individual hypotheses are rejected at $\alpha = 5\%$ level of significance. All comparisons among the distributions from the control group and any active treatment group show an significant effect. In particular, all pairwise comparisons among the distributions from the active treatments do not demonstrate any significance at 5% level. Summing up we can find a significant effect of the active treatment, but we have also seen that this significance is only due to differences among the distributions from the reaction times between the control and active treatment groups, respectively.

6. Conclusions and future work

The R package **nparcomp** implements a broad range of rank-based nonparametric methods for multiple comparisons. The single step procedures provide local test decisions in terms of multiplicity adjusted p-values and simultaneous confidence intervals. They further allow for user-defined contrasts. In particular, paradox results in terms of Efron's paradox dice cannot occur (Thangavelu and Brunner 2006). The derivation of stepwise confidence intervals, however, is part of future research. A notable novel feature of **nparcomp** is that it can easily be extended to the anaylsis of factorial designs. We plan to update the package **nparcomp** on a regular basis with new nonparametric statistical procedures available for multiple comparisons. In addition, we plan to undertake a major update of the code and release **nparcomp** in the S4 style.

All implemented methods in the package are based on asymptotic results on the distribution of rank statistics. We plan to derive adequate resampling- and permutation based methods to approximate the distribution of the statistics for very small sample sizes. For example, optimal subset procedures and weighted methods controlling the familywise error rate are proposed by Finos and Salmaso (2005), Finos and Salmaso (2006) and Finos and Salmaso (2007). Permutation tests for umbrella alternatives and multivariate problems are considered in Basso and Salmaso (2011) and Pesarin and Salmaso (2012).

Acknowledgments

This work was supported by the German Science Foundation project DFG-BR 655/16-1 and DFG-HO-1687/9-1.

References

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

Basso D, Salmaso L (2011). "A Permutation Test for Umbrella Alternatives." *Statistics and Computing*, **21**, 45–54.

Bretz F (2006). "An Extension of the Williams Trend Test to General Unbalanced Linear Models." *Computational Statistics & Data Analysis*, **50**(7), 1735–1748.

Bretz F, Genz A, Hothorn LA (2001). "On the Numerical Availability of Multiple Comparison Procedures." *Biometrical Journal*, **43**(5), 645–656.

Bretz F, Hothorn T, Westfall P (2010). *Multiple Comparisons Using R*. Chapman and Hall, London.

Browne RH (2010). "The t-Test p-Value and Its Relationship to the Effect Size and $P(X > Y)$." *The American Statistician*, **64**(1), 30–33.

Brunner E, Munzel U (2000). "The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation." *Biometrical Journal*, **1**(1), 17–21.

Brunner E, Munzel U (2002). *Nichtparametrische Datenanalyse*. Springer-Verlag, New York.

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

Campbell G, Skillings JH (1985). "Nonparametric Stepwise Multiple Comparison Procedures." *Journal of the American Statistical Association*, **80**, 998–1003.

Dunnett CW (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control." *Journal of the American Statistical Association*, **50**(272), 1096–1121.

Finos L, Salmaso L (2004). "Nonparametric Multi-Focus Analysis for Categorical Variables." *Communication in Statistics – Theory and Methods*, **33**, 1931–1941.

Finos L, Salmaso L (2005). "A New Nonparametric Approach for Multiplicity Control: Optimal Subset Procedures." *Computational Statistics*, **20**, 643–654.

Finos L, Salmaso L (2006). "Weighted Methods Controlling the Multiplicity when the Number of Variables Is Much Higher than the Number of Observations." *Journal of Nonparametric Statistics*, **18**, 245–261.

Finos L, Salmaso L (2007). "FDR- and FWE-controlling Methods Using Data-Driven Weights." *Journal of Statistical Planning and Inference*, **137**, 3859–3870.

Gao X, Alvo M, Chen J, Li G (2008). "Nonparametric Multiple Comparison Procedures for Unbalanced One-Way Factorial Designs." *Journal of Statistical Planning and Inference*, **138**(8), 2574–2591.

Genz A, Bretz F (2009). *Computation of Multivariate Normal and t Probabilities*. Springer-Verlag, New York.

Hirotsu C (1997). "Two-Way Change-Point Model and Its Application." *Australian Journal of Statistics*, **39**(2), 205–218.

Hochberg Y (1986). "A Sharper Bonferroni Procedure for Multiple Tests of Significance." *Biometrika*, **75**, 800–802.

Hothorn T, Bretz F, Westfall P (2008). "Simultaneous Inference in General Parametric Models." *Biometrical Journal*, **50**(3), 346–363.

Konietschke F (2009). *Simultane Konfidenzintervalle für Nichtparametrische Relative Kontrasteffekte*. Ph.D. thesis, Georg-August Universität Göttingen.

Konietschke F (2015). ***nparcomp**: Perform Multiple Comparisons and Compute Simultaneous Confidence Intervals for the Nonparametric Relative Contrast Effects*. R package version 2.6, URL http://CRAN.R-project.org/package=nparcomp.

Konietschke F, Bösiger S, Brunner E, Hothorn LA (2013). "Are Multiple Contrast Tests Superior to the ANOVA?" *The International Journal of Biostatistics*, **9**, 1–11.

Konietschke F, Hothorn LA (2012). "Evaluation of Toxicological Studies Using a Nonparametric Shirley-type Trend Test for Comparing Several Dose Levels with a Control Group." *Statistics in Biopharmaceutical Research*, **4**, 14–27.

Konietschke F, Hothorn LA, Brunner E (2012a). "Rank-Based Multiple Test Procedures and Simultaneous Confidence Intervals." *Electronic Journal of Statistics*, **6**, 1–8.

Konietschke F, Libiger O, Hothorn LA (2012b). "Nonparametric Evaluation of Quantitative Traits in Population-Based Association Studies when the Genetic Model is Unknown." *PLoS ONE*, **7**(2), e31242. doi:10.1371/journal.pone.0031242.

Marozzi M (2009). "Some Notes on the Location-Scale Cucconi Test." *Journal of Nonparametric Statistics*, **21**(1–2), 629–647.

Marozzi M (2013). "Nonparametric Simultaneous Tests for Location and Scale Testing: A Comparison of Several Methods." *Communications in Statistics – Simulation and Computation*, **42**, 1298–1317.

Munzel U, Hothorn LA (2001). "A Unified Approach to Simultaneous Rank Test Procedures in the Unbalanced One-Way Layout." *Biometrical Journal*, **43**(5), 553–569.

Neubert K, Brunner E (2007). "A Studentized Permutation Test for the Nonparametric Behrens-Fisher Problem." *Computational Statistics & Data Analysis*, **51**, 5192 – 5204.

Pesarin F, Salmaso L (2010). *Permutation Tests for Complex Data*. John Wiley & Sons, Chichester.

Pesarin F, Salmaso L (2012). "A Review and Some New Results on Permutation Testing for Multivariate Problems." *Statistics and Computing*, **22**, 639–646.

Ryu E (2009). "Simultaneous Confidence Intervals Using Ordinal Effect Measures for Ordered Categorical Outcomes." *Statistics in Medicine*, **28**(25), 3179–3188.

Ryu E, Agresti A (2008). "Modeling and Inference for an Ordinal Effect Size Measure." *Statistics in Medicine*, **27**(25), 1703 – 1717.

Shirley E (1977). "Nonparametric Equivalent of Williams Test for Contrasting Increasing Dose Levels of a Treatment." *Biometrics*, **33**(2), 386–389.

Steel RDG (1960). "A Rank Sum Test for Comparing All Pairs of Treatments." *Technometrics*, **2**, 197 – 207.

Thangavelu K, Brunner E (2006). "Wilcoxon Mann-Whitney Test and Efron's Paradox Dice." *Journal of Statistical Planning and Inference*, **137**, 720 – 737.

Tukey JW (1953). "The Problem of Multiple Comparisons." Unpublished manuscript reprinted in *The Collected Works of John W. Tukey, Volume VIII*, Chapman and Hall, London.

Williams DA (1972). "The Comparison of Several Dose Levels with a Zero Dose Control." *Biometrics*, **28**(2), 519–531.

Affiliation:

Frank Konietschke
Department of Mathematical Sciences
The University of Texas at Dallas
800 W Campbell Road
75080 Richardson, TX, United States of America
E-mail: fxk141230@utdallas.edu

Marius Placzek
Department of Medical Statistics
University of Göttingen
D-37073 Göttingen, Germany
E-mail: marius.placzek@stud.uni-goettingen.de

Frank Schaarschmidt, Ludwig A. Hothorn
Institute of Biostatistics
Leibniz University of Hannover
D-30419 Hannover, Germany
E-mail: schaarschmidt@biostat.uni-hannover.de, hothorn@biostat.uni-hannover.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 64, Issue 9 *Submitted:* 2012-11-08
March 2015 *Accepted:* 2014-06-10