Brunner, Edgar; Konietschke, Frank; Bathke, Arne C.; Pauly, Markus

Ranks and Pseudo-ranks—Surprising Results of Certain Rank Tests in Unbalanced Designs

International Statistical Review

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Brunner, Edgar; Konietschke, Frank; Bathke, Arne C.; Pauly, Markus (2021) : Ranks and Pseudo-ranks—Surprising Results of Certain Rank Tests in Unbalanced Designs, International Statistical Review, ISSN 1751-5823, Wiley, Hoboken, NJ, Vol. 89, Iss. 2, pp. 349-366, https://doi.org/10.1111/insr.12418

International Statistical Review

International Statistical Review (2021), 89, 2, 349–366 doi:10.1111/insr.12418

Ranks and Pseudo-ranks—Surprising Results of Certain Rank Tests in Unbalanced Designs

Edgar Brunner[1], Frank Konietschke[2], Arne C. Bathke[3] and Markus Pauly[4]

[1]*Department of Medical Statistics, University of Göttingen, Göttingen, Germany*
E-mail: ebrunne1@gwdg.de

[2]*Institute of Biometry and Clinical Epidemiology, Charité, Berlin, Germany*

[3]*Intelligent Data Analytics (IDA) Lab, Salzburg, Austria*

[4]*Faculty of Statistics, TU Dortmund University, Dortmund, Germany*

Summary

Rank-based inference methods are applied in various disciplines, typically when procedures relying on standard normal theory are not justifiable. Various specific rank-based methods have been developed for two and more samples and also for general factorial designs (e.g. Kruskal–Wallis test or Akritas–Arnold–Brunner test). It is the aim of the present paper (1) to demonstrate that traditional rank procedures for several samples or general factorial designs may lead to surprising results in case of unequal sample sizes as compared with equal sample sizes, (2) to explain why this is the case and (3) to provide a way to overcome these disadvantages. Theoretical investigations show that the surprising results can be explained by considering the non-centralities of the test statistics, which may be non-zero for the usual rank-based procedures in case of unequal sample sizes, while they may be equal to 0 in case of equal sample sizes. A simple solution is to consider unweighted relative effects instead of weighted relative effects. The former effects are estimated by means of the so-called pseudo-ranks, while the usual ranks naturally lead to the latter effects. A real data example illustrates the practical meaning of the theoretical discussions.

Key words: Rank statistic; pseudo-rank statistic; Kruskal–Wallis test; Hettmansperger–Norton test; Akritas–Arnold–Brunner test; unweighted relative effect; weighted relative effect.

1 Introduction

When the assumptions of classical parametric inference methods are not met, the usual recommendation is to apply non-parametric rank-based tests. Here, the Wilcoxon–Mann–Whitney and Kruskal & Wallis (1952) tests are among the most commonly applied rank procedures, often utilised as replacements for the unpaired two-sample *t*-test and the one-way analysis of variance (ANOVA), respectively. Other common rank methods include the Hettmansperger &

Norton (1987) test for ordered alternatives and the procedures by Akritas *et al.* (1997) for two-way or higher-way designs. In statistical practice, these procedures are usually appreciated as robust and powerful inference tools when standard assumptions are not fulfilled. For example, Whitley & Ball (2002) conclude that 'Nonparametric [rank-based] methods require no or very limited assumptions to be made about the format of the data, and they may, therefore, be preferable when the assumptions required for parametric methods are not valid'.

These descriptions are slightly overoptimistic because rank-based methods also rely on certain assumptions. Furthermore, they are based on different effects than contrasts of means.

In case of doubt, it is nevertheless expected that rank procedures are more robust and lead to more reliable results than their parametric counterparts. While this is true for deviations from normality, and while by now it is clear that ordinal data in general should rather be analysed using adequate rank-based methods than using normal theory procedures, we illustrate in various instances that non-parametric rank tests for more than two samples possess one noteworthy weakness. Namely, they are generally non-robust against changes from balanced to unbalanced designs. In particular, keeping the set of distributions fixed, we provide paradigms under which commonly used rank tests surprisingly yield *completely opposite* test decisions when rearranging group sample sizes. These examples need not be in general artificially generated to obtain surprising results but even include homoscedastic normal models in two-way layouts. The practical meaning of the theoretical considerations is demonstrated by a real data example in Section 6.

In order to comprehensively answer the question whether rank procedures can adequately handle designs for more than two groups, we carefully analyse the underlying non-parametric effects of the respective rank procedures. It should be noted, however, that rank procedures address different statistical models than simply considering the usual contrasts of means in shift models. Moreover, many rank procedures for more than two groups are based on weighted relative effects where each distribution is compared with the weighted mean of all distributions in the experiment, and the weights are the relative sample sizes. These weighted relative effects are 'estimated' by the means of the ranks where the expression 'estimated' is a slight abuse of the terminology because the weighted relative effects are not fixed model quantities. Basically, we use three different types of rankings: (1) the usual (overall) ranks, (2) the pseudo-ranks and (3) the so-called *internal ranks*, which may not be confused with the *within-block ranks*, which are used in the Friedman (1937) test for block designs. The usual ranks (simply called 'ranks' throughout this paper) are used to evaluate the weighted relative effects, while the pseudo-ranks are used to estimate the unweighted relative effects. The internal ranks are only used to estimate the variances of the rank statistics by computing the so-called *placements* (Orban & Wolfe, 1980,1982). These three types of ranks are defined in (1), (19) and (21) below. For more details, we refer to the textbook by Brunner *et al.* (2019), definition 2.20. The weighted relative effects are evaluated by functions of the rank means. Throughout this paper, however, they will also be called 'estimators' for convenience, but having in mind this slight abuse of terminology. The simple idea to overcome the problems of rank procedures in case of unequal sample sizes is to define unweighted relative effects where each distribution is compared with the unweighted mean of all distributions in the experiment. These unweighted relative effects basically have the same intuitive interpretation as the weighted relative effects. They are estimated by the so-called *pseudo-ranks*, which have already been considered by Kulle (1999), Gao & Alvo (2005a, 2005b), and in more detail by Thangavelu & Brunner (2007), and by Brunner *et al.* (2017). However, it should be noted that the motivation in these referenced articles was different and that their authors had not been aware of the striking properties that may arise when using rank tests in case of unequal sample sizes. These unexpected paradigms only appear in case of unequal sample sizes, and the surprising results

mentioned earlier cannot occur with rank procedures, which are defined only for equal samples sizes as, for example, the Friedman (1937) test and the Kepner & Robinson (1988) test. Pseudo-ranks are easy to compute (Happ *et al.*, 2019), share the same advantageous properties of ranks and lead to reliable and robust inference procedures for a variety of factorial designs. Moreover, we can even obtain confidence intervals for (contrasts of) easy to interpret reasonable non-parametric effects, namely, the unweighted relative effects. This way, they also contribute to resolve the widespread (but wrong) perception that 'nonparametric [rank-based] methods are geared towards hypothesis testing rather than estimation of effects' (Whitley & Ball, 2002). To the best of our knowledge, these unweighted relative effects have been first mentioned in the literature by Brunner & Puri (2001, Sections 1.3.1 and 3.2).

The paper is organised as follows. Notations are introduced in Section 2. Then in Section 3, some surprising results are presented in the one-way layout for the Kruskal–Wallis test and for the Hettmansperger–Norton trend test by means of certain tricky (non-transitive) dice. In the two-way layout, a more striking result for the Akritas–Arnold–Brunner test in a simple 2×2 design is presented in Section 4 using a homoscedastic normal shift model. The theoretical background of the unexpected results is discussed in Section 5, and a solution of the problem by using the unweighted relative effects, which ultimately lead to the pseudo-ranks, is investigated in detail. Finally, the computation of confidence intervals for the purely non-parametric effects is briefly discussed in Section 5.2. A real data example demonstrating the issues discussed earlier is provided in Section 6. The paper closes with some discussions and conclusions regarding an adequate application of rank procedures.

2 Statistical Model and Notations

We consider rank tests in factorial designs for $d > 2$ samples of $N = \sum_{i=1}^{d} n_i$ independent observations $X_{ik} \sim F_i = \frac{1}{2}[F_i^- + F_i^+]$, $i = 1, \ldots, d$, $k = 1, \ldots, n_i$. Let R_{ik} denote the rank of X_{ik} among all N observations, then a rank test is usually based on the means $\overline{R}_{i\cdot}$ of the ranks

$$R_{ik} = \frac{1}{2} + \sum_{r=1}^{d}\sum_{\ell=1}^{n_r} c(X_{ik} - X_{r\ell}), \tag{1}$$

where $c(u) = 0, 1/2, 1$ for $u <, =$ or >0, respectively, denotes the count function. To determine the rank tests' consistency region, we have to find the theoretical quantities estimated by the rank means. To this end, note that $\frac{1}{n_r}\sum_{\ell=1}^{n_r} c(X_{ik} - X_{r\ell}) = \widehat{F}_r(X_{ik})$ is the value of the empirical distribution function of the observations $X_{r1}, \ldots, X_{rn_r}$ within sample r at the random point X_{ik}. Thus, we have $n_r \widehat{F}_r(X_{ik}) = \sum_{\ell=1}^{n_r} c(X_{ik} - X_{r\ell})$, and R_{ik} in (1) can be rewritten as (see, e.g. Akritas *et al.*, 1997, Formula (14))

$$R_{ik} = \frac{1}{2} + \sum_{r=1}^{d} n_r \widehat{F}_r(X_{ik}) = \frac{1}{2} + N\widehat{H}(X_{ik}), \tag{2}$$

where $\widehat{H}(x) = \frac{1}{N}\sum_{r=1}^{d} n_r \widehat{F}_r(x)$ denotes the weighted mean of the empirical distribution functions $\widehat{F}_r(x)$. Because $E\left[\widehat{F}_r(X_{ik})\right] = \int F_r dF_i$, $i = 1, \ldots, r; k = 1, \ldots, n_i$ (see, e.g. Lemma 7.4 in Brunner *et al.*, 2019), it follows that

$$E(R_{ik}) = \frac{1}{2} + N p_i, \text{ where } p_i = \int H dF_i \tag{3}$$

and $H = \frac{1}{N}\sum_{r=1}^{d} n_r F_r$ denotes the weighted mean of the distributions $F_1, \ldots, F_d$. Thus, $E\left(\overline{R}_{i\cdot}\right) = N p_i + \frac{1}{2}$ is linearly related to p_i in (3).

An unbiased and consistent estimator of p_i is obtained by the simple plug-in estimator

$$\widehat{p}_i = \int \widehat{H}\, d\,\widehat{F}_i = \frac{1}{N}\left(\overline{R}_{i\cdot} - \frac{1}{2}\right), \tag{4}$$

where $\widehat{F}_i$ denotes the (normalised) empirical distribution of $X_{i1}, \ldots, X_{in_i}, i = 1, \ldots, d$, and $\widehat{H} = \frac{1}{N}\sum_{r=1}^{d} n_r \widehat{F}_r$ their weighted mean and $\overline{R}_{i\cdot}$ is the mean of the ranks R_{ik}. This means that ranks are simple and intuitive tools to estimate the quantity $p_i = P(Z < X_{i1}) + \frac{1}{2}P(Z = X_{i1})$, which is, easily interpreted, the probability that a randomly selected observation Z from the weighted mean distribution H is smaller than a randomly selected observation X_{i1} from the distribution F_i plus $\frac{1}{2}$ times the probability that both observations are equal. Because p_i depends on the weights n_i/N, it is strictly speaking not an effect, that is, a fixed model quantity by which hypotheses could be formulated and for which confidence intervals could be given. Nevertheless, we follow the historical notion and call p_i a *(weighted) relative effect* (see, e.g. Brunner & Puri, 2001; Thangavelu & Brunner, 2007) because it measures a difference of the distribution F_i with respect to the weighted mean distribution H.

As a consequence, we do not consider rank tests as procedures where simply the observations are replaced by their ranks. Instead, we consider ranks as quantities used to obtain unbiased and consistent estimators of purely non-parametric effects defined by the distributions F_i and the relative sample sizes n_i/N. Considering linear functions or quadratic forms of these non-parametric weighted relative effects enables a different look at rank procedures, which come out in a natural way by this approach. To this end, the weighted relative effects p_i are arranged in the vector $\boldsymbol{p} = (p_1, \ldots, p_d)' = \int H d\boldsymbol{F}$, where $\boldsymbol{F} = (F_1, \ldots, F_d)'$ denotes the vector of the distributions and the estimators $\widehat{p}_1, \ldots, \widehat{p}_d$ are arranged in the vector

$$\widehat{\boldsymbol{p}} = \int \widehat{H}\, d\,\widehat{\boldsymbol{F}} = \frac{1}{N}\left(\overline{\boldsymbol{R}}_{\cdot} - \frac{1}{2}\mathbf{1}_d\right), \tag{5}$$

where $\widehat{\boldsymbol{F}} = (\widehat{F}_1, \ldots, \widehat{F}_d)'$ is the vector of the empirical distributions, $\overline{\boldsymbol{R}}_{\cdot} = (\overline{R}_{1\cdot}, \ldots, \overline{R}_{d\cdot})'$ the vector of the rank means $\overline{R}_{i\cdot}$, and $\mathbf{1}_d = (1, \ldots, 1)'_{d\times 1}$ denotes the vector of 1s.

2.1 Relation to the Mann–Whitney Effect for $d = 2$ Samples

The non-parametric weighted relative effects p_i for $d > 2$ samples are generalisations of the Mann–Whitney effect for $d = 2$ samples $X_{ik} \sim F_i, i = 1, 2; k = 1, \ldots, n_i$. In the case of two independent random variables $X_1 \sim F_1$ and $X_2 \sim F_2$, Birnbaum & Klose (1957) had called the function $L(t) = F_2[F_1^{-1}(t)]$ the 'relative distribution function' of X_1 and X_2, assuming continuous distributions. Thus, its expectation

$$\int_0^1 t\, dL(t) = \int_{-\infty}^{\infty} F_1(s) d\, F_2(s) = P(X_1 < X_2)$$

is called a 'relative effect' with an obvious adaptation of the notation. Depending on the context, it is also known as probabilistic index (e.g. Acion *et al.*, 2006; Thas *et al.*, 2012), Mann–Whitney & Wilcoxon effect (e.g. Janssen, 1999; Chung & Romano, 2016; Dobler *et al.*, 2020) or stress–strength characteristic (e.g. Kotz *et al.*, 2003).

In the case of several samples, the weighted relative effect p_i is a linear combination of the pairwise effects $w_{ri} = \int F_r d F_i$. In vector notation, the quantities p_i in (3) are written as

$$\boldsymbol{p} = \int H d\boldsymbol{F} = \boldsymbol{W}'\boldsymbol{n} = \begin{pmatrix} w_{11} & \dots & w_{d1} \\ \vdots & \ddots & \vdots \\ w_{1d} & \dots & w_{dd} \end{pmatrix} \cdot \begin{pmatrix} n_1/N \\ \vdots \\ n_d/N \end{pmatrix} = \begin{pmatrix} p_1 \\ \vdots \\ p_d \end{pmatrix}. \tag{6}$$

Here, $\boldsymbol{F} = (F_1, \dots, F_d)'$ denotes the vector of distribution functions, and

$$\boldsymbol{W} = \int \boldsymbol{F} d\boldsymbol{F}' = \begin{pmatrix} w_{11} & \dots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{d1} & \dots & w_{dd} \end{pmatrix} \tag{7}$$

is the matrix of the pairwise effects w_{ri}. Note that $w_{ii} = \frac{1}{2}$ and $w_{ir} = 1 - w_{ri}$, which follows from integration by parts.

The earlier decomposition of the distribution H as a weighted sum of the distributions F_i shows that the effects underlying the known rank procedures for $d > 2$ samples can be represented as linear combinations of the pairwise Mann–Whitney effects $w_{ri} = \int F_r d F_i$. In what follows, the dependence on the relative sample sizes n_i/N of the consistency regions of rank tests for $d > 2$ will be demonstrated by means of this decomposition in the one-way layout in (13) and in the two-way layout in (17).

In the following sections, we demonstrate that for $d \geq 3$ groups, rank tests may lead to striking results in case of unequal sample sizes. In particular, for factorial designs involving two or more factors, the non-parametric main effects and interactions (defined by the weighted relative effects $p_{ij} = \int H d F_{ij}$) may be severely biased.

3 Surprising Results in the One-way Layout

To demonstrate some surprising results of rank tests for $d \geq 3$ samples in the one-way layout, we consider the vector $\boldsymbol{p} = \int H d\boldsymbol{F}$ of the weighted relative effects p_i. Let $\widehat{\boldsymbol{p}} = \int \widehat{H} d\widehat{\boldsymbol{F}}$ denote the plug-in estimator of $\widehat{\boldsymbol{p}}$ defined in (5). In order to detect whether the p_i are different, we study the asymptotic distribution of $\sqrt{N}\widehat{\boldsymbol{p}}$, which is obtained from the asymptotic equivalence theorem

$$\sqrt{N}(\widehat{\boldsymbol{p}} - \boldsymbol{p}) \doteqdot \sqrt{N}\left[\overline{\boldsymbol{Y}}_\cdot + \overline{\boldsymbol{Z}}_\cdot - 2\boldsymbol{p}\right], \tag{8}$$

if $N/n_i \leq N_0 < \infty$. This condition is assumed throughout the paper. For details, we refer to Akritas *et al.* (1997), Brunner & Puri (2001, 2002) or Brunner *et al.* (2019) for different generalisations.

In (8), the symbol $\doteqdot$ denotes asymptotic equivalence, while the mean vectors $\overline{\boldsymbol{Y}}_\cdot = \int H d\widehat{\boldsymbol{F}}$ and $\overline{\boldsymbol{Z}}_\cdot = \int \widehat{H} d\boldsymbol{F}$ have expectations $E(\overline{\boldsymbol{Y}}_\cdot) = E(\overline{\boldsymbol{Z}}_\cdot) = \boldsymbol{p}$. It follows from the central limit theorem that $\sqrt{N}\left[\overline{\boldsymbol{Y}}_\cdot + \overline{\boldsymbol{Z}}_\cdot - 2\boldsymbol{p}\right]$ has, asymptotically, a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}_N$, which has a quite involved structure (for details, see Brunner *et al.*, 2017).

However, the covariance matrix simplifies dramatically when testing the hypothesis $H_0^F(\boldsymbol{T})$: $\boldsymbol{T}\boldsymbol{F} = \boldsymbol{0}$, where $\boldsymbol{T}$ is an appropriate contrast matrix, which can be assumed to be a projection matrix without loss of generality. Under $H_0^F(\boldsymbol{T})$, this follows from

$$\sqrt{N}\boldsymbol{T}(\widehat{\boldsymbol{p}} - \boldsymbol{p}) \doteqdot \sqrt{N}\boldsymbol{T}\left[\overline{\boldsymbol{Y}}_\cdot + \overline{\boldsymbol{Z}}_\cdot - 2\boldsymbol{p}\right] = \sqrt{N}\boldsymbol{T}\overline{\boldsymbol{Y}}_\cdot.$$

because $\boldsymbol{T}\overline{\boldsymbol{Z}}. = \boldsymbol{T}\int \widehat{H}\, d\boldsymbol{F} = \int \widehat{H}\, d(\boldsymbol{TF}) = \boldsymbol{0}$ and $\boldsymbol{Tp} = \boldsymbol{T}\int H d\boldsymbol{F} = \int H d(\boldsymbol{TF}) = \boldsymbol{0}$ (see Akritas & Arnold, 1994). For example, to test the hypothesis $H_0^F : F_1 = \ldots = F_d$ that all distributions are equal in the one-way layout, usually the centring matrix $\boldsymbol{T}_d = \boldsymbol{I}_d - \frac{1}{d}\boldsymbol{J}_d$ is chosen as a contrast matrix. Here, $\boldsymbol{I}_d$ denotes the d-dimensional unit matrix and $\boldsymbol{J}_d = \boldsymbol{1}_d \boldsymbol{1}_d'$ the $d \times d$-dimensional matrix of 1s. The asymptotic distribution of $\sqrt{N}\boldsymbol{T}_d(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ is the multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{T}_d \boldsymbol{\Sigma}_N \boldsymbol{T}_d$, where $\boldsymbol{\Sigma}_N = \mathrm{Cov}\left(\sqrt{N}\left[\overline{\boldsymbol{Y}}. + \overline{\boldsymbol{Z}}.\right]\right)$. Then it follows from (8) that in general,

$$\sqrt{N}\boldsymbol{T}_d\widehat{\boldsymbol{p}} \doteqdot \sqrt{N}\boldsymbol{T}_d\left[\overline{\boldsymbol{Y}}. + \overline{\boldsymbol{Z}}. - 2\boldsymbol{p}\right] + \sqrt{N}\boldsymbol{T}_d\boldsymbol{p}. \tag{9}$$

Obviously, the multivariate distribution is shifted by $\sqrt{N}\boldsymbol{T}_d\boldsymbol{p}$ from the origin $\boldsymbol{0}$. Therefore, $\boldsymbol{T}_d\boldsymbol{p}$ is the particular quantity defining the rank tests' consistency region and will be called 'multivariate non-centrality'. A corresponding 'univariate non-centrality' may be quantified by the quadratic form $c_p = \boldsymbol{p}'\boldsymbol{T}_d\boldsymbol{p}$. In particular, we have $c_p = 0$ iff $\boldsymbol{T}_d\boldsymbol{p} = \boldsymbol{0}$. The actual (multivariate) shift of the distribution, depending on the total sample size N, is $\sqrt{N}\boldsymbol{T}_d\boldsymbol{p}$, and the corresponding univariate non-centrality (depending on N) is then given by $N \cdot c_p$. From these considerations, it follows that $N \cdot c_p \to \infty$ as $N \to \infty$ if $\boldsymbol{T}_d\boldsymbol{p} \neq \boldsymbol{0}$. This defines the consistency region of a test based on $\sqrt{N}\boldsymbol{T}_d\widehat{\boldsymbol{p}}$, which leads to the Kruskal–Wallis test. For the general technical derivation, we refer to Brunner & Puri (2001, section 1.6.1).

In the succeeding text, we will demonstrate that for the same set of distributions $\boldsymbol{F} = (F_1, \ldots, F_d)'$, the non-centrality $c_p = \boldsymbol{p}'\boldsymbol{T}_d\boldsymbol{p}$ may be 0 in case of equal sample sizes, while c_p may be unequal to 0 in case of unequal sample sizes. This means that for equal sample sizes, $\boldsymbol{F}$ does not belong to the consistency region $\boldsymbol{T}_d\boldsymbol{p} \neq \boldsymbol{0}$, while for unequal sample sizes, it may be contained in the consistency region. Under the strong non-parametric hypothesis formulated in terms of the distribution functions $H_0^F : \boldsymbol{T}_d\boldsymbol{F} = \boldsymbol{0}$, it follows that $\boldsymbol{T}_d\boldsymbol{p} = \boldsymbol{0}$. If, however, the strong non-parametric hypothesis H_0^F is not true, then the non-centrality $c_p = \boldsymbol{p}'\boldsymbol{T}_d\boldsymbol{p}$ may be 0 or unequal to 0 for the same set of distributions $F_1, \ldots, F_d$, because c_p depends on the relative samples sizes $n_1/N, \ldots, n_d/N$. As a consequence, existing rank tests for H_0^F may reject the null hypothesis simply due to a reallocation of the designs from balanced to unbalanced (while all other parts of the settings remained fixed).

Some well-known tests in the one-way layout, which have this undesirable property, are, for example, the Kruskal & Wallis (1952) test and the Hettmansperger & Norton (1987) trend test.

As an example, consider the case of $d = 3$ distributions given by the probability mass functions $f_1(x) = \frac{1}{6}$ if $x \in \{9, 16, 17, 20, 21, 22\}$, $f_2(x) = \frac{1}{6}$ if $x \in \{13, 14, 15, 18, 19, 26\}$, $f_3(x) = \frac{1}{6}$ if $x \in \{10, 11, 12, 23, 24, 25\}$ and $f_1(x) = f_2(x) = f_3(x) = 0$ otherwise. These discrete distributions are derived from some tricky dice (see, e.g. Peterson, 2002). For the distribution functions $F_i(x)$ defined by $f_i(x)$, $i = 1, 2, 3$ earlier, it is easily seen that

$$w_{21} = P(X_2 < X_1) = \int F_2 d\, F_1 = 7/12, \tag{10}$$

$$w_{13} = P(X_1 < X_3) = \int F_1 d\, F_3 = 7/12, \tag{11}$$

$$w_{32} = P(X_3 < X_2) = \int F_3 d\, F_2 = 7/12. \tag{12}$$

Table 1. *Ratios of relative sample sizes n_i/N, weighted relative effects p_i, and the non-centralities for the example of the tricky dice where $w = 7/12$ and the distributions F_1, F_2 and F_3 are fixed.*

Setting	n_1/N	n_2/N	n_3/N	p_1	p_2	p_3	$\overline{p}_\cdot$	c_p
(A)	1/3	1/3	1/3	0.5	0.5	0.5	0.5	0
(B)	2/3	1/12	1/4	0.4861	0.4653	0.5486	0.5	0.00376
(C)	1/4	2/3	1/12	0.5486	0.4861	0.4653	0.5	0.00376

Thus, $w_{21} = w_{13} = w_{32} = w$, and the vector of the weighted relative effects p_i is given by

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \boldsymbol{W}'\boldsymbol{n} = \frac{1}{N} \begin{pmatrix} \frac{1}{2}n_1 + n_3 + (n_2 - n_3)w \\ n_1 + \frac{1}{2}n_2 + (n_3 - n_1)w \\ n_2 + \frac{1}{2}n_3 + (n_1 - n_2)w \end{pmatrix}. \tag{13}$$

Equation (13) demonstrates that for equal samples sizes, the vector $\boldsymbol{p}$ of the weighted relative effects is equal to the constant $\frac{1}{2}\mathbf{1}_3$ while it depends in general on the pairwise differences of the relative samples sizes n_i/N. It is easily seen that in this example,

$$F_1 = F_2 = F_3 \Rightarrow\ p_1 = p_2 = p_3 \iff \begin{cases} w = \frac{1}{2} \text{ or} \\ n_1 = n_2 = n_3 \equiv n. \end{cases} \tag{14}$$

This means that $p_1 = p_2 = p_3$ holds if $w \neq \frac{1}{2}$ while the samples sizes are equal, but if $w \neq \frac{1}{2}$ in case of unequal sample sizes, then the weighted relative effects p_1, p_2 and p_3 are different. In the latter case, $\boldsymbol{T}_d\boldsymbol{p}$ is contained in the consistency region, while in the former case, it is not contained in the consistency region—for the same set of distributions!

The weighted relative effects p_i of the three discrete distributions and the resulting non-centralities c_p are listed in Table 1 for equal and some different unequal sample sizes.

Because for unequal sample sizes one obtains $c_p \neq 0$ if $w \neq \frac{1}{2}$, it is only a question of choosing the total sample size N large enough to reject the hypothesis $H_0^F : F_1 = F_2 = F_3$ by the Kruskal–Wallis test with a probability arbitrary close to 1 (based on $\chi^2_{2;1-\alpha}$ as the critical value). In case of equal sample sizes for $N \to \infty$, the probability of rejecting the hypothesis remains constant equal to α^* (close to α) because in this case, $c_p = 0$. It may be noted that in general, $\alpha^* \neq \alpha$ because the variance estimator of the Kruskal–Wallis statistic is computed under the strong hypothesis $H_0^F : F_1 = F_2 = F_3$, which is not true in general. Thus, the scaling is not correct, and the Kruskal Wallis test has a slightly different type I error α^*.

For the Hettmansperger–Norton trend test, the situation gets worse because for different ratios of sample sizes, the non-parametric effects p_1, p_2 and p_3 may change their order. In setting (B) in Table 1, we have $p_2 < p_1 < p_3$, while in setting (C), we have $p_3 < p_2 < p_1$. Now consider the non-centrality of the Hettmansperger–Norton trend test, which is a linear rank test. Let $\boldsymbol{c} = (c_1, \ldots, c_d)'$ denote a vector reflecting the conjectured pattern. Then it follows from (9) that

$$\sqrt{N}\boldsymbol{c}'\boldsymbol{T}_d\widehat{\boldsymbol{p}} \doteq \sqrt{N}\boldsymbol{c}'\boldsymbol{T}_d\left[\overline{\boldsymbol{Y}}_\cdot + \overline{\boldsymbol{Z}}_\cdot - 2\boldsymbol{p}\right] + \sqrt{N}\boldsymbol{c}'\boldsymbol{T}_d\boldsymbol{p}. \tag{15}$$

The quantity $c_p^{HN} = \boldsymbol{c}'\boldsymbol{T}_d\boldsymbol{p}$ is a univariate non-centrality, and if $\boldsymbol{T}_d\boldsymbol{p} = \boldsymbol{0}$, then it follows that $c_p^{HN} = \boldsymbol{c}'\boldsymbol{T}_d\boldsymbol{p} = 0$. If, however, $\boldsymbol{T}_d\boldsymbol{p} \neq \boldsymbol{0}$, then $c_p^{HN} < 0$ indicates a decreasing trend and $c_p^{HN} > 0$ an increasing trend. In the earlier discussed example, we obtain for setting (B) and for

a conjectured pattern of $\boldsymbol{c} = (1, 2, 3)'$ the non-centrality $c_p^{HN} = \sum_{i=1}^{3} c_i \left(p_i - \frac{1}{2}\right) = 1/16 > 0$, indicating an increasing trend. For setting (C), however, we obtain $c_p^{HN} = -1/12$, indicating a decreasing trend. In case of setting (A) (equal sample sizes), $c_p^{HN} = 0$ because $p_1 = p_2 = p_3 = \frac{1}{2}$ and thus indicating no trend. Again, it is a question of choosing the total sample size N large enough to obtain the decision of a significantly decreasing trend for the first setting (B) of unequal sample sizes and for the second setting (C) the decision of a significantly increasing trend with a probability arbitrary close to 1 for the same distributions F_1, F_2 and F_3. These results are contradicting.

4 Surprising Results in the Two-way Layout

In the previous section, surprising decisions by rank tests in case of unequal sample sizes were demonstrated for the one-way layout using large sample sizes and particular configurations of distributions leading to non-transitive decisions. In this section, we will show that in two-way layouts, similar unexpected results may already occur in simple homoscedastic normal shift models. To this end, we consider the simple 2×2 design with two crossed factors A and B, each with two levels $i = 1, 2$ for A and $j = 1, 2$ for B. The observations $X_{ijk} \sim F_{ij}$, $k = 1, \ldots, n_{ij}$, are assumed to be independent.

The hypotheses of no non-parametric effects in terms of the distribution functions $F_{ij}(x)$ are expressed as $H_0^F(A) : F_{11} + F_{12} - F_{21} - F_{22} = 0$ (no main effect of factor A), $H_0^F(B) : F_{11} - F_{12} + F_{21} - F_{22} = 0$ (no main effect of factor B) and $H_0^F(AB) : F_{11} - F_{12} - F_{21} + F_{22} = 0$ (no interaction AB), where in all three cases, 0 denotes a function, which is identical 0. We nevertheless would like to point out that all of these null hypotheses do not imply exchangeability of the rank vector in general. Thus, the corresponding rank tests cannot be directly performed as permutation tests. For a detailed discussion, we refer to Akritas *et al.* (1997) as well as to Umlauft *et al.* (2017).

Let $\boldsymbol{F} = (F_{11}, F_{12}, F_{21}, F_{22})'$ denote the vector of these distribution functions. Then the three hypotheses formulated earlier can be written in matrix notation as $H_0^F(\boldsymbol{c}) : \boldsymbol{c}'\boldsymbol{F} = 0$, where $\boldsymbol{c} = \boldsymbol{c}_A = (1, 1, -1, -1)'$ for the main effect of factor A, $\boldsymbol{c} = \boldsymbol{c}_B = (1, -1, 1, -1)'$ for the main effect of factor B and $\boldsymbol{c} = \boldsymbol{c}_{AB} = (1, -1, -1, 1)'$ for the interaction AB. For testing these hypotheses, Akritas *et al.* (1997) derived rank procedures based on the statistic

$$L_N(\boldsymbol{c}) = \sqrt{N}\boldsymbol{c}'\widehat{\boldsymbol{p}} = \frac{1}{\sqrt{N}}\boldsymbol{c}'\overline{\boldsymbol{R}}., \tag{16}$$

where $\overline{\boldsymbol{R}}. = (\overline{R}_{11\cdot}, \overline{R}_{12\cdot}, \overline{R}_{21\cdot}, \overline{R}_{22\cdot})'$ denotes the vector of the rank means $\overline{R}_{ij\cdot}$ for the four samples. We note that this statistic is based on the non-parametric functional $\boldsymbol{c}' \int H d\boldsymbol{F}$. Under the null hypothesis $H_0^F(\boldsymbol{c}) : \boldsymbol{c}'\boldsymbol{F} = \boldsymbol{0}$, the statistic $L_N(\boldsymbol{c})$ in (16) has, asymptotically, a normal distribution with mean 0 and variance $\sigma_0^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{N}{n_{ij}}\sigma_{ij}^2$, where the unknown variances σ_{ij}^2 are consistently estimated under $H_0^F(\boldsymbol{c})$ by using the ranks R_{ijk} of the observations X_{ijk}.

To demonstrate a surprising result, we assume that the observations X_{ijk} are coming from normal distributions $N(\mu_{ij}, \tau^2)$ with equal standard deviations $\tau^2 = 0.6$ and with expectations $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})' = (4, 5, 5, 6)'$. From the viewpoint of linear models, there is a main effect A of $\boldsymbol{c}_A'\boldsymbol{\mu} = \mu_{11} + \mu_{12} - \mu_{21} - \mu_{22} = -2$, a main effect B of $\boldsymbol{c}_B'\boldsymbol{\mu} = \mu_{11} - \mu_{12} + \mu_{21} - \mu_{22} = -2$ but no AB interaction because $\boldsymbol{c}_{AB}'\boldsymbol{\mu} = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$. Because this

is a homoscedastic linear model, the classical ANOVA should reject the hypotheses $H_0^{\mu}(\boldsymbol{c}_A) : \boldsymbol{c}_A'\boldsymbol{\mu} = 0$ and $H_0^{\mu}(\boldsymbol{c}_B) : \boldsymbol{c}_B'\boldsymbol{\mu} = 0$ with a high probability if the total sample size is large enough, while the hypothesis $H_0^{\mu}(\boldsymbol{c}_{AB}) : \boldsymbol{c}_{AB}'\boldsymbol{\mu} = 0$ of no interaction should only be rejected with the preselected type I error probability α. The non-centralities are given by $c_\mu(A) = \boldsymbol{c}_A'\boldsymbol{\mu} = -2$, $c_\mu(B) = \boldsymbol{c}_B'\boldsymbol{\mu} = -2$ and $c_\mu(AB) = \boldsymbol{c}_{AB}'\boldsymbol{\mu} = 0$. In this example, $F_{11} = N(4, \tau^2)$, $F_{12} = F_{21} = N(5, \tau^2)$ and $F_{22} = N(6, \tau^2)$, where $\tau^2 = 0.6$. Thus, for the non-parametric non-centrality, one obtains (the details are given in Section 2.1 of the Supporting Information)

$$\begin{aligned} c_p(AB) = \boldsymbol{c}_{AB}'\boldsymbol{p} &= p_{11} - p_{12} - p_{21} + p_{22} \\ &= \frac{n_{11} - n_{22}}{N}\left(\frac{1}{2} - 2w + v\right) \approx -0.18 \cdot (n_{11} - n_{22})/N, \end{aligned} \tag{17}$$

where $w = \int F_{11} dF_{12} = \Phi\left(1/\left(\sqrt{1.2}\right)\right) = 0.823$ and $v = \int F_{11} dF_{22} = \Phi\left(2/\sqrt{1.2}\right) \approx 0.966$. Thus, if $n_{11} \neq n_{22}$, it follows that the non-parametric non-centrality $c_p(AB) \neq 0$, while it is equal to 0 if $n_{11} = n_{22}$. In contrast to that, the non-centrality $c_\mu(AB)$ for the ANOVA is equal to 0 in both cases because $c_\mu(AB)$ does not depend on the sample sizes. This means that in the case of equal sample sizes $n_{11} = n_{22}$, both the ANOVA and the rank test in (16) should only reject the hypothesis of no interaction with probability close to the nominal level α. For unequal sample sizes $n_{11} \neq n_{22}$, however, the rank test will reject the hypothesis of no interaction with a probability arbitrary close to 1 if the total sample size N is large enough, whereas the ANOVA will again have a rejection probability close to the nominal α.

In order to investigate the dependence of this rejection probability on the difference $n_{11} - n_{22}$ while the total sample size N is fixed, we have performed a simulation study in this setting, and the results are reported in Section 2.2 of the Supporting Information. For example, the sample sizes $n_{11} = 800$, $n_{12} = 20$, $n_{21} = 30$ and $n_{22} = 10$ lead to $c_p(AB) = -4.65$, and the rejection probability was about 64%, while for equal sample sizes $n_{ij} \equiv n = 215$, the non-centrality is zero, and the rejection probability was about 3.5%.

On the surface, the difference of the two non-centralities $c_\mu(AB)$ and $c_p(AB)$ in the unbalanced case could be explained by the fact that the non-parametric hypothesis $H_0^F(AB)$ and the parametric hypothesis $H_0^{\mu}(AB)$ are not identical and that this particular configuration of normal distributions falls into the region where $H_0^{\mu}(AB)$ is true, but $H_0^F(AB)$ is not. It is surprising, however, that this explanation does not hold for the balanced case. This calls for an explanation.

5 Explanation of the Surprising Results

The simple reason for the surprising results is the fact that even when all distribution functions underlying the observations are fixed, the consistency region $c_p = \boldsymbol{p}'\boldsymbol{T}_d\boldsymbol{p} \neq 0$ of a rank test based on $\widehat{\boldsymbol{p}}$ is not fixed in general because $\boldsymbol{p} = \int H d\boldsymbol{F}$ depends on the relative sample sizes n_i/N. Thus, in general, $\boldsymbol{p}$ is not a fixed model quantity by which hypotheses could be formulated or for which confidence intervals could be reasonably computed.

The details shall be demonstrated by means of Figure 1, which shows an example of the consistency regions of the Kruskal–Wallis test and the ANOVA for equal and unequal sample sizes. The point $\boldsymbol{f} = (f_1, f_2, f_3)$ represents the vector of three different discrete distributions considered in Section 3 in the three-dimensional space of distributions. Note that these distributions are different, and thus, $H_0^F : F_1 = F_2 = F_3$ is not true while $H_0^{\mu} : \mu_1 = \mu_2 = \mu_3$ is true because the expectations μ_i are equal to 9/2, $i = 1, \ldots, 3$. The area outside the solid

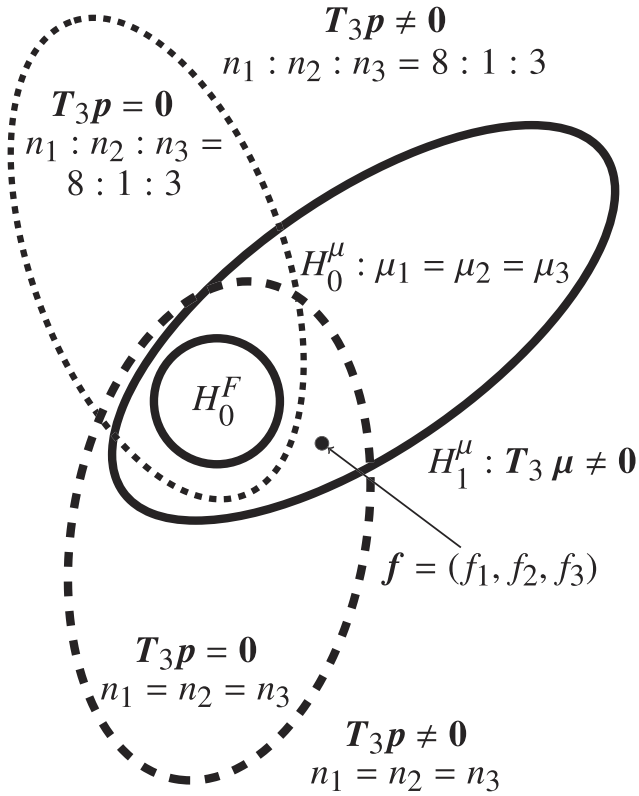

Figure 1. *Graphical representation of the hypotheses and the consistency regions of the analysis of variance and the Kruskal–Wallis test for the three discrete distributions f_1, f_2 and f_3 from Section 3. The solid region is fixed, and the dashed region refers to the situation of equal sample sizes, while the dotted region varies with the ratios of the sample sizes. The fixed point $\boldsymbol{f} = (f_1, f_2, f_3)$ may or may not be contained in the consistency region of the Kruskal–Wallis test $c_p^{KW} = \boldsymbol{p}'\boldsymbol{T}_3\boldsymbol{p} \neq 0$ only by changing the ratios of the sample sizes.*

ellipse is the consistency region $c_\mu = \boldsymbol{\mu}'\boldsymbol{T}_3\,\boldsymbol{\mu} \neq 0$ of the ANOVA, the area outside the dashed ellipse is the consistency region $c_p \neq 0$ of the Kruskal–Wallis test involving equal sample sizes, while the area outside the dotted ellipse is the consistency region $c_p \neq 0$ of the Kruskal–Wallis test involving unequal sample sizes $n_1 : n_2 : n_3 = 8{:}1{:}3$. In the example of the three discrete distributions in Section 3, the non-centrality of the ANOVA does not depend on the relative sample sizes and is given by $c_\mu = \boldsymbol{\mu}'\boldsymbol{T}_3\,\boldsymbol{\mu} = 0$ for the distributions in this example, so that the point $\boldsymbol{f}$ is not contained in the consistency region of the ANOVA. The non-centrality of the Kruskal–Wallis test involving equal sample sizes is given by $c_p = \boldsymbol{p}'\boldsymbol{T}_3\boldsymbol{p} = 0$ because $p_1 = p_2 = p_3 = 1/2$ and is thus contained within the dashed ellipse. For unequal sample sizes (setting (B) in Table 1), however, the non-centrality of the Kruskal–Wallis test is given by $c_p = 0.00376 \neq 0$ because $p_1 = 0.4861$, $p_2 = 0.4653$ and $p_3 = 0.5486$. This means that the consistency region of the Kruskal–Wallis test is not a fixed region: it may or may not contain the fixed point $\boldsymbol{f}$ corresponding to the distributions given in this example only by varying the ratio of the sample sizes.

In the one-way layout, one needs to employ crossing distribution functions leading to non-transitive decisions in order to demonstrate the strange phenomenon that non-centralities and subsequently test decisions may depend on the ratio of sample sizes. However, in the two-way layout, this phenomenon can already appear with shifted homoscedastic normal distributions as

demonstrated by the example in Section 4. The reasons are the same as in the one-way layout, namely, that (1) there is a gap between the strong hypothesis H_0^F and the consistency region of a rank test and (2) this consistency region is not fixed; instead, it depends on the ratio of the sample sizes. This unfortunately also provides a possibility to manipulations.

5.1 Unweighted Effects and Pseudo-Ranks

With the foregoing considerations in mind, it appears reasonable to define different nonparametric effects, which are fixed model quantities not depending on sample sizes. To this end, let $G = \frac{1}{d}\sum_{r=1}^{d} F_r$ denote the unweighted mean distribution, and let $\psi_i = \int G\, dF_i$. This nonparametric effect ψ_i can be interpreted as the probability that a randomly drawn observation from the mean distribution function $G(x)$ is smaller than a randomly drawn observation from the distribution function $F_i(x)$ (plus $\frac{1}{2}$ times the probability that it is equal). Thus, the quantity ψ_i measures an effect of the distribution F_i with respect to the unweighted mean distribution G and is therefore a 'fixed relative effect'. It can be estimated consistently by the simple plug-in estimator

$$\widehat{\psi}_i = \int \widehat{G}\, d\widehat{F}_i = \frac{1}{N}\left(\overline{R}_{i\cdot}^{\psi} - \frac{1}{2}\right), \tag{18}$$

where $\widehat{G} = \frac{1}{d}\sum_{r=1}^{d}\widehat{F}_r$ denotes the unweighted mean of the empirical distributions $\widehat{F}_1, \ldots, \widehat{F}_d$, and $\overline{R}_{i\cdot}^{\psi} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ik}^{\psi}$ the mean of the so-called *pseudo-ranks*

$$\text{ps-rank}(X_{ik}) = R_{ik}^{\psi} = \frac{1}{2} + N\widehat{G}(X_{ik}) = \frac{1}{2} + \frac{N}{d}\sum_{r=1}^{d}\frac{1}{n_r}\sum_{\ell=1}^{n_r} c(X_{ik} - X_{r\ell}). \tag{19}$$

The only difference of Equations (19) and (2) is that $\widehat{H}(X_{ik})$ in (2) is replaced with $\widehat{G}(X_{ik})$ in (19). Comparing (18) and (19), it is easily seen that the expectation of the pseudo-rank mean $\overline{R}_{i\cdot}^{\psi}$ is linearly related to the unweighted relative effect ψ_i, namely, $E\left(\overline{R}_{i\cdot}^{\psi}\right) = N\psi_i + \frac{1}{2}$. From a formal and technical viewpoint, this means that ranks are replaced by the corresponding pseudo-ranks.

Finally, the unweighted relative effects ψ_i are arranged in the vector $\psi = \int G\, d\boldsymbol{F}$, and the estimators $\widehat{\psi}_i$ are arranged in the vector

$$\widehat{\boldsymbol{\psi}} = \left(\widehat{\psi}_1, \ldots, \widehat{\psi}_d\right)' = \int \widehat{G}\, d\widehat{\boldsymbol{F}} = \frac{1}{N}\left(\overline{\boldsymbol{R}}_{\cdot}^{\psi} - \frac{1}{2}\mathbf{1}_d\right), \tag{20}$$

where $\overline{\boldsymbol{R}}_{\cdot}^{\psi} = (\overline{R}_{1\cdot}^{\psi}, \ldots, \overline{R}_{d\cdot}^{\psi})'$ is the vector of the pseudo-rank means $\overline{R}_{i\cdot}^{\psi}$.

The pseudo-ranks R_{ik}^{ψ} are also order preserving and invariant under strictly monotone transformations (see Lemma 1 of the Supporting Information). Both the ranks and the pseudo-ranks are obtained from ranking the observations albeit resulting in different quantities with similar properties. Therefore, the quantities R_{ik}^{ψ} are referred to as 'pseudo-ranks'. We use this denomination also in grateful memory of Steven Arnold († 2014) who first mentioned this name in a discussion about rank statistics at Penn State University about a quarter century ago.

Furthermore the pseudo-ranks can be represented as linear combinations of internal ranks within sample i,

$$R_{ik}^{(i)} = \frac{1}{2} + \sum_{\ell=1}^{n_r} c(X_{ik} - X_{i\ell}), \tag{21}$$

and all pairwise ranks $R_{ik}^{(ir)}$ within samples $i \neq r = 1, \ldots, d$, namely,

$$R_{ik}^{\psi} = \frac{1}{2} + \frac{N}{d}\left[\sum_{r\neq i}^{d}\frac{1}{n_r}\left(R_{ik}^{(ir)} - R_{ik}^{(i)}\right) + \frac{1}{n_i}\left(R_{ik}^{(i)} - \frac{1}{2}\right)\right]. \tag{22}$$

It should be noted that the relation in (22) is only a technical relation of the pseudo-ranks to some other rankings (pairwise rankings and internal ranks). The definition of a pseudo-rank R_{ik}^{ψ} in (19) does neither require pairwise rankings nor internal ranks. Some more details along with some particular properties of the pseudo-ranks are given in Sections 1.1 and 1.2 of the Supporting Information.

5.2 Consistency Regions of Pseudo-rank Procedures

Here, we demonstrate that replacing the ranks R_{ik} with the pseudo-ranks R_{ik}^{ψ} leads to procedures that do not have the surprising results discussed in Sections 3 and 4. The main reason is that pseudo-rank procedures are based on the (unweighted) relative effects ψ_i, which are fixed model quantities, by which hypotheses can be formulated and for which confidence intervals can be derived. Thus, it appears reasonable to formulate non-parametric hypotheses by these quantities in the same way as for the distribution functions. Let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)' = \int G d\boldsymbol{F}$, then a general non-parametric hypothesis about the fixed relative effects $\boldsymbol{\psi}$ can be expressed as

$$H_0^{\psi}(\boldsymbol{T}) : \boldsymbol{T}\boldsymbol{\psi} = \boldsymbol{0}, \tag{23}$$

where $\boldsymbol{T}$ denotes an appropriate hypothesis matrix and $\boldsymbol{0} = (0, \ldots, 0)'$. An asymptotic equivalence theorem—similar to that for the rank procedures in (8)—has been established by Konietschke *et al.* (2012) in the context of confidence intervals and by Brunner *et al.* (2017) in the general case. This theorem states that

$$\sqrt{N}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \doteq \sqrt{N}\left[\overline{\boldsymbol{Y}}_{\cdot}^{\psi} + \overline{\boldsymbol{Z}}_{\cdot}^{\psi} - 2\boldsymbol{\psi}\right], \tag{24}$$

where $\overline{\boldsymbol{Y}}_{\cdot}^{\psi} = \int G d\widehat{\boldsymbol{F}}$ and $\overline{\boldsymbol{Z}}_{\cdot}^{\psi} = \int \widehat{G} d\boldsymbol{F}$ are vectors of means of random vectors with expectations $E(\overline{\boldsymbol{Y}}_{\cdot}^{\psi}) = E(\overline{\boldsymbol{Z}}_{\cdot}^{\psi}) = \boldsymbol{\psi}$. It follows from the central limit theorem that $\sqrt{N}\left[\overline{\boldsymbol{Y}}_{\cdot}^{\psi} + \overline{\boldsymbol{Z}}_{\cdot}^{\psi} - 2\boldsymbol{\psi}\right]$ has, asymptotically, a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}_N$, which has a quite involved structure (for details, see Konietschke *et al.*, 2012; Brunner *et al.*, 2017). The consistency region of a test for the hypothesis $H_0^{\psi}(\boldsymbol{T})$: $\boldsymbol{T}\boldsymbol{\psi} = \boldsymbol{0}$ follows from (24) and

$$\sqrt{N}\boldsymbol{T}\widehat{\boldsymbol{\psi}} \doteq \sqrt{N}\boldsymbol{T}\left[\overline{\boldsymbol{Y}}_{\cdot}^{\psi} + \overline{\boldsymbol{Z}}_{\cdot}^{\psi} - 2\boldsymbol{\psi}\right] + \sqrt{N}\boldsymbol{T}\boldsymbol{\psi}. \tag{25}$$

Because under $H_0^{\psi}(\boldsymbol{T})$ the quantity $\sqrt{N}\boldsymbol{T}\left[\overline{\boldsymbol{Y}}_{\cdot}^{\psi} + \overline{\boldsymbol{Z}}_{\cdot}^{\psi}\right]$ has, asymptotically, a multivariate normal distribution with expectation $\boldsymbol{0}$ and covariance matrix $\boldsymbol{T}\boldsymbol{V}_N\boldsymbol{T}$, the consistency region is given by $\boldsymbol{T}\boldsymbol{\psi} \neq \boldsymbol{0}$ or equivalently by $c_{\psi} = \boldsymbol{\psi}'\boldsymbol{T}\boldsymbol{\psi} = 0$. The details about the covariance matrix $\boldsymbol{V}_N = \mathrm{Cov}\left(\sqrt{N}\boldsymbol{T}\left[\overline{\boldsymbol{Y}}_{\cdot}^{\psi} + \overline{\boldsymbol{Z}}_{\cdot}^{\psi}\right]\right)$ and a consistent estimator $\widehat{\boldsymbol{V}}_N$ using pairwise rankings are derived in Konietschke *et al.* (2012) and Brunner *et al.* (2017) but are not in the focus of this paper.

In the succeeding text, we examine the behaviour of pseudo-rank procedures in the situations where the use of rank tests led to the surprising results by considering the consistency region in (25).

1. For testing the hypothesis $H_0^F : F_1 = F_2 = F_3 \iff \boldsymbol{T}_3\boldsymbol{F} = \boldsymbol{0}$ in case of the tricky dice in the example in Section 3, one obtains for the non-centrality of the Kruskal–Wallis statistic the value $c_{\psi}^{KW} = \psi' \boldsymbol{T}_3 \psi = 0$ if the weighted relative effects p_i are replaced by the unweighted relative effects ψ_i. This means that the ranks R_{ik} are replaced with the pseudo-ranks R_{ik}^{ψ} for estimating the unweighted instead of the weighted effects. In this case, it follows from (10), (11), (12) and (13) that $\boldsymbol{\psi} = \frac{1}{2}\mathbf{1}_3$ and $c_{\psi}^{KW} = \boldsymbol{\psi}' \boldsymbol{T}_3 \boldsymbol{\psi} = \frac{1}{2}\mathbf{1}_3' \boldsymbol{T}_3 \frac{1}{2}\mathbf{1}_3 = 0$ by noting that $\boldsymbol{T}_3$ is a contrast matrix, and thus, $\boldsymbol{T}_3\mathbf{1}_3 = \boldsymbol{0}$.
2. When substituting ranks with pseudo-ranks in the example in Section 3, then the unweighted effects ψ_i are estimated and the non-centrality for the Hettmansperger–Norton trend test becomes $c_{\psi}^{HN} = \boldsymbol{c}' \boldsymbol{T}_3 \boldsymbol{\psi} = 0$ for all trend alternatives $\boldsymbol{c} = (c_1, c_2, c_3)'$.
3. In the two-way layout, we reconsider the example of the four shifted normal distributions (for details, we refer to Equation 6 of the Supporting Information). We obtain for the unweighted relative effects ψ_{ij}

$$\boldsymbol{\psi} = \boldsymbol{W}' \cdot \frac{1}{4}\mathbf{1}_4 = \frac{1}{4}\begin{pmatrix} 7/2 - 2w - v \\ 2 \\ 2 \\ 1/2 + 2w + v \end{pmatrix}$$

and the non-centrality of the statistic for the interaction

$$c_{\psi}(AB) = \psi_{11} - \psi_{12} - \psi_{21} + \psi_{22} = 0. \tag{26}$$

In all these cases, surprising results obtained by changing the ratios of the sample sizes cannot occur because the non-centralities c_{ψ}^{KW}, c_{ψ}^{HN} and $c_{\psi}(AB)$ are equal to 0 for all constellations of the relative sample sizes. In case of equal sample sizes $n_i \equiv n$, $i = 1, \ldots, d$, we do not obtain surprising results by rank tests because in this case, ranks and pseudo-ranks coincide, $R_{ik} = R_{ik}^{\psi}$.

It is clear by these considerations that such unexpected results for the well-known Friedman (1937) test and the Kepner & Robinson (1988) test cannot occur because by design, the sample sizes (i.e. the number of blocks) are equal for all treatments.

In case of $d = 2$ samples, it is easily seen that $p_2 - p_1 = \psi_2 - \psi_1 = p = \int F_1 d F_2$, which does not depend on sample sizes. Thus, surprising results for rank-based tests—such as presented in the previous sections—can only occur for $d \geq 3$ samples.

Confidence intervals for linear combinations $\boldsymbol{c}'\boldsymbol{\psi}$ of the fixed relative effects $\boldsymbol{\psi}$ are easily obtained from (24). The details are given in Konietschke *et al.* (2012) and Brunner *et al.* (2017). For multivariate designs, see Dobler *et al*., (2020) and Umlauft *et al.* (2019).

6 Real Data Example of a Subgroup Analysis

Here we discuss a register study for patients with multiple sclerosis (MS). It should be investigated on possible pitfalls that may occur in the baseline comparability of MS cohorts from different types of centres. The centres and the distribution of relapsing-remitting (RR) and primary progressive (PP) disease courses may be heterogeneous, so that differences in baseline covariables must be accounted for in the analysis of effects, for example, in symptomatic

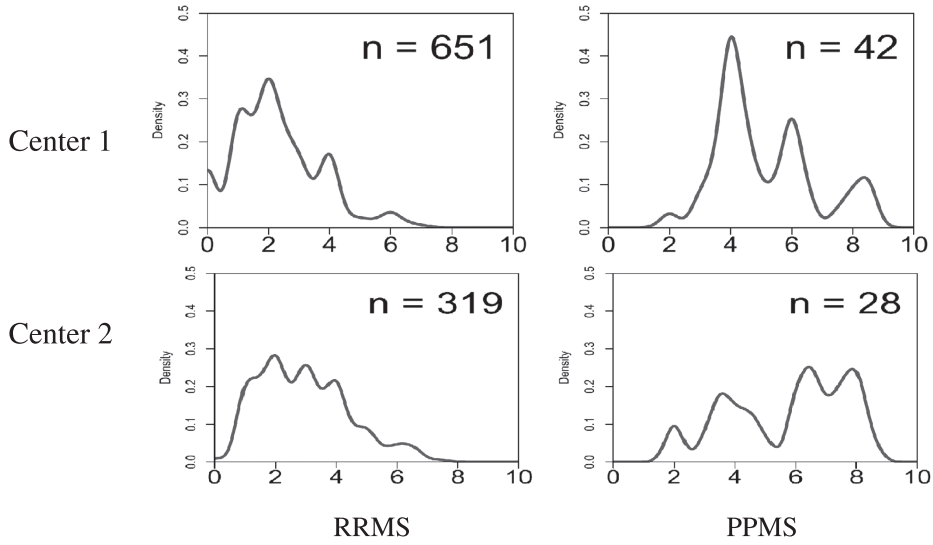

Figure 2. *Densities of the expanded disability scale scores within the two centres for the multiple sclerosis subgroups relapsing-remitting multiple sclerosis (left) and primary progressive multiple sclerosis (right).*

treatment, behavioural interventions or socio-demographics. The present study focuses on the assessment of differences in the expanded disability scale (EDSS) as a measure of disability in MS patients. The EDSS is a bounded outcome score (0–10) in steps of 0.5. The differences are compared between two different types of centres and their interaction with the disease courses of RRMS and PPMS. For details, we refer to Rommer *et al.* (2018).

The analyses of the register study for patients with MS are performed using three different models—just for comparison. In model 1, the effects are described by the weighted relative effects $p_{ij} = \int H d F_{ij}$, which are estimated by means of the ranks, while in model 2, the effects are described by the unweighted relative effects $\psi_{ij} = \int G d F_{ij}$, which are estimated by means of the pseudo-ranks. These two models consider the EDSS as ordinal data. Model 3, however, considers this scale as metric data. An appropriate analysis may then be an asymptotic ANOVA because the total sample size is quite large. The asymptotic ANOVA uses the quantiles of the χ^2 distribution as critical values instead of the F distribution.

The frequency distributions smoothed by a kernel density estimator are displayed in Figure 2.

The estimated weighted and unweighted relative effects as well as the mean-based effects for both types of MS and their differences δ_i^p, δ_i^ψ and δ_i^μ within both centres are listed in Table 2.

The results of the tests for the main effects of the centres and of the type of the MS as well as for the interaction are displayed in Table 3.

Obviously, the interaction $L_N^p(\boldsymbol{c}_{AB})$ is biased by the ratio of the samples sizes appearing in the weighted relative effects. This becomes clear from Equation (2) and from Equation (20) because both quantities $\boldsymbol{p}$ and $\boldsymbol{\psi}$ are based on the same pairwise effects $w_{rs} = \int F_r d F_s$. The only difference is that for $\boldsymbol{p}$, the linear combinations of the w_{rs} are weighted by the samples sizes n_i/N, while this is not the case for $\boldsymbol{\psi}$. It is noteworthy that the effects in this example described by the means and by the unweighted relative effects are quite close for the two centres. In any case, the effects described by the unweighted relative effects cannot be changed by different ratios of the sample sizes—similar as for the means.

Table 2. *Estimators of the relative effects $\widehat{p}_{ij}$ and $\widehat{\psi}_{ij}$ as well as the means $\widehat{X}_{ij\cdot}$ of the two subgroups RRMS and PPMS within the two centres.*

	Estimated effects and differences								
	$\widehat{p}_{ij}$ (weighted)			$\widehat{\psi}_{ij}$ (unweighted)			$\widehat{X}_{ij\cdot}$ (means)		
Centre i	Type RR	Type PP	δ_i^p	Type RR	Type PP	δ_i^ψ	Type RR	Type PP	δ_i^μ
$i=1$	0.431	0.859	0.428	0.260	0.662	0.402	2.23	5.10	2.87
$i=2$	0.560	0.875	0.315	0.358	0.720	0.362	2.94	5.73	2.79

The differences $\delta_i^\psi = \widehat{\psi}_{i2} - \widehat{\psi}_{i1}$, $i = 1, 2$, of the unweighted relative effects and of the means $\delta_i^\mu = \widehat{X}_{i2\cdot} - \widehat{X}_{i1\cdot}$ are quite close when compared between the two centres, while the differences $\delta_i^p = \widehat{p}_{i2} - \widehat{p}_{i1}$ of the weighted relative effects are much more different between the two centres.
PP, primary progressive multiple sclerosis; RR, relapsing-remitting multiple sclerosis.

Table 3. *Analyses of the EDSS of the MS study by the statistics $\boldsymbol{L}_N^p(\boldsymbol{c})$ in Equation (16) based on the weighted relative effects p_{ij} using the ranks, $\boldsymbol{L}_N^\psi(\boldsymbol{c})$ based on the unweighted relative effects $\boldsymbol{\psi}_{ij}$ using the pseudo-ranks and based on $\boldsymbol{L}_N^\mu$ using the means μ_{ij} by an asymptotic ANOVA. The p-values listed in the table are two sided.*

	Ranks		Pseudo-ranks		Means	
Effect	L_N^p	p-value	L_N^ψ	p-value	L_N^μ	p-value
Centre	3.45	0.0009	2.85	0.0061	2.92	0.0035
MS type	17.88	$<10^{-4}$	13.95	$<10^{-4}$	12.34	$<10^{-4}$
Centre × type	2.71	0.0084	0.72	0.4752	0.14	0.8880

ANOVA, analysis of variance; EDSS, expanded disability scale; MS, multiple sclerosis.

7 Conclusions and Discussion

We have demonstrated that in designs involving more than two samples, certain rank tests may lead to surprising results. The reason for this is that in factorial designs, the expectations of the rank means depend on the relative sample sizes n_i/N in the design. These expectations have been denoted as *relative effects* (Brunner & Puri, 2001), *Mann–Whitney–Wilcoxon* effects (e.g. Janssen, 1999; Chung & Romano, 2016), *stress–strength characteristic* (e.g. Kotz *et al.*, 2003) or *probabilistic index* (e.g. Acion *et al.*, 2006). Indeed, for $d \geq 3$ samples, these quantities are *weighted effects* that depend on the relative sample sizes and are thus not fixed model quantities. Moreover, we have shown that the consistency regions of rank tests based on these quantities also depend on the relative sample sizes. This means that the consistency regions are not fixed and may or may not contain a fixed set of distribution functions depending on the ratio of the sample sizes. This is ultimately the reason why surprising results with rank tests may happen in case of unequal sample sizes.

There are two exceptions:

(1) In the case of two samples the Mann–Whitney effect, $\int F_1 \, d\, F_2$ does not depend on sample sizes. Thus, unequal sample sizes in case of $d = 2$ samples cannot lead to striking results.
(2) Whenever the sample sizes are identical in the different treatment groups by design. For example, the Friedman (1937) test or the Kepner & Robinson (1988) test were developed for blocks of equal lengths. Thus, the number of blocks is the same for each treatment, and unexpected results cannot occur. This is, however, not true for modifications of these procedures. For example, if unequal numbers of replications within the different blocks

and treatments are performed or if missing values occur in an unbalanced pattern (see also Ramosaj *et al.*, 2018, for additional difficulties that may appear in this context), then also surprising results for these tests are possible.

An obvious solution of the problem with rank tests in case of unequal sample sizes and more than two treatments is to consider *unweighted relative effects*. Here, also each distribution function is compared with the mean distribution function—as in the case of the rank tests—but the difference is that the unweighted mean of the distribution functions is used instead of the weighted mean. These unweighted relative effects have basically the same intuitive and simple interpretations as the weighted relative effects and have similar statistical properties. They can be estimated consistently by the so-called pseudo-ranks, which are obtained as simple plug-in estimators from the definition of the unweighted relative effects. A similar asymptotic equivalence theorem (Thangavelu & Brunner, 2007; Brunner *et al.*, 2017) as for the rank statistics enables the investigation of the consistency regions of tests based on the unweighted relative effects. Obviously, these consistency regions are based upon fixed quantities and thus cannot vary with changing sample sizes. Moreover, confidence intervals can be given (Konietschke *et al.*, 2012) so that these non-parametric procedures not only are appropriate for hypothesis testing but also enable a quantitative description of the magnitude of the non-parametric effect and its variability in the experiment.

Technically, these methods can be obtained from existing rank procedures by replacing ranks with pseudo-ranks. Such a limiting look at the pseudo-rank procedures would be quite superficial hiding the real ideas of the derivation. However, a careful examination of the situations where the surprising results with the rank tests appeared showed that using pseudo-rank procedures solves these problems. This was theoretically demonstrated by the examination of the non-centralities of the pseudo-rank procedures. Finally, the real data example in Section 6 demonstrates the meaning of the theoretical considerations. This example may also serve as a cautionary note when using non-parametric methods in subgroup analysis.

In case of factorial designs with independent observations, pseudo-rank procedures are already implemented in the R package *rankFD* by choosing the option for *unweighted* effects (Konietschke *et al.*, 2019). A quick method for the computation of the pseudo-ranks is provided in the R package *pseudorank* (Happ *et al.*, 2019).

Acknowledgements

The work of Arne Bathke and Markus Pauly was supported by the Austrian Science Fund [Fonds zur Förderung der wissenschaftlichen Forschung (FWF)] (grant number I 2697-N31) and the German Research Foundation [Deutsche Forschungsgemeinschaft (DFG)] (grant number PA 2409/4-1) both within a joined D-A-CH Lead Agency Project. The work of Frank Konietschke was supported by the DFG (grant number KO 4680/3-2). The authors are grateful to MS Forschungs- und Projektentwicklungs-gGmbH (a subsidiary of the German MS Foundations of the German Multiple Sclerosis Society DMSG, Bundesverband e.V.) for providing the data example from the German MS Register. The nationwide registry for persons with MS in Germany was initiated in 2001 and collects clinical data on disease characteristics, treatment and healthcare use. We also would like to thank David Ellenberger for assistance in data processing and Julian Streitberger (Salzburg) for assistance with designing Figure 1. Particular thanks go to the editors and three anonymous referees for their helpful and constructive comments, which greatly improved the manuscript. Open access funding enabled and organized by Projekt DEAL.

References

Acion, L., Peterson, J., Temple, S. & Arndt, S. (2006). Probabilistic index: An intuitive nonparametric approach to measuring the size of treatment effects. *Stat. Med.*, **25**, 591–602.

Akritas, M.G. & Arnold, S.F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *J. Am. Stat. Assoc.*, **89**, 336–343.

Akritas, M.G., Arnold, S.F. & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *J. Am. Stat. Assoc.*, **92**, 258–265.

Birnbaum, Z.W. & Klose, O.M. (1957). Bounds for the variance of the Mann–Whitney statistic. *Ann. Math. Stat.*, **28**, 933–945.

Brunner, E., Bathke, A.C. & Konietschke, F. (2019). *Rank- and Pseudo-rank Procedures in Factorial Designs—Using R and SAS—Independent Observations Springer Series in Statistics*. Springer: Heidelberg.

Brunner, E., Konietschke, F., Pauly, M. & Puri, M.L. (2017). Rank-based procedures in factorial designs: hypotheses about nonparametric treatment effects. *J. R. Stat. Soc. Ser. B*, **79**, 1463–1485.

Brunner, E. & Puri, M.L. (2001). Nonparametric methods in factorial designs. *Stat. Pap.*, **42**, 1–52.

Brunner, E. & Puri, M.L. (2002). A class of rank-score tests in factorial designs. *J. Stat. Plann. Inference*, **103**, 331–360.

Chung, E.Y. & Romano, J.P. (2016). Asymptotically valid and exact permutation tests based on two-sample *U*-statistics. *J. Stat. Plann. Inference*, **168**, 97–105.

Dobler, D., Friedrich, S. & Pauly, M. (2020). Nonparametric MANOVA in meaningful effects. *Ann Inst Stat Math*, **72** 997–1022. https://doi.org/10.1007/s10463-019-00717-3

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, **32**, 675–701.

Gao, X. & Alvo, M. (2005a). A nonparametric test for interaction in two-way layouts. *Can. J. Stat.*, **33**, 529–543.

Gao, X. & Alvo, M. (2005b). A unified nonparametric approach for unbalanced factorial designs. *J. Am. Stat. Assoc.*, **100**, 926–941.

Happ, M., Zimmermann, G., Brunner, E. & Bathke, A. (2019). Pseudo-ranks: How to calculate them efficiently in R. *J. Stat. Softw.*, **95**(1), 1–22. https://doi.org/10.18637/jss.v095.c01.

Hettmansperger, T.P. & Norton, R.M. (1987). Tests for patterned alternatives in *k*-sample problems. *J. Am. Stat. Assoc.*, **82**, 292–299.

Janssen, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *J. Stat. Plann. Inference*, **81**, 71–93.

Kepner, J.L. & Robinson, D.H. (1988). Nonparametric methods for detecting treatment effects in repeated-measures designs. *J. Am. Stat. Assoc.*, **83**, 456–461.

Konietschke, F., Friedrich, S., Brunner, E. & Pauly, M. (2019). RankFD: rank-based tests for general factorial designs. R package version 0.0.3.

Konietschke, F., Hothorn, L.A. & Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electron. J. Stat.*, **6**, 737–758. https://doi.org/10.1214/12-EJS691

Kotz, S., Lumelskii, Y. & Pensky, M. (2003). *The Stress–Strength Model and Its Generalizations*. World Scientific Publishing: New Jersey.

Kruskal, W.H. & Wallis, W.A. (1952). The use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.*, **47**, 583–621.

Kulle, B. (1999). Nichtparametrisches Behrens-Fisher-Problem im Mehrstichprobenfall. Diploma Thesis, University of Göttingen.

Orban, J. & Wolfe, D.A. (1980). Distribution-free partially sequential placement procedures. *Communications in Statistics, Ser. A*, **9**, 883–904.

Orban, J. & Wolfe, D.A. (1982). A class of distribution-free two-sample tests based on placements. *J. Am. Stat. Assoc.*, **77**, 666–672.

Peterson, I. (2002). Tricky dice revisited. *Sci. News*, 161. https://www.sciencenews.org/article/tricky-dice-revisited

Ramosaj, B., Amro, L. & Pauly, M. (2018). A cautionary tale on using imputation methods for inference in matched pairs design. arXiv:1806.06551 [stat.AP].

Rommer, P.S., Eichstädt, K., Ellenberger, D., Flachenecker, P., Friede, T., Haas, J., Kleinschnitz, C., Pöhlau, D., Rienhoff, O., Stahmann, A. & Zettl, U.K. (2018). Symptomatology and symptomatic treatment in multiple sclerosis: Results from a nationwide MS registry. *Mult. Scler. J.*, **25**, 1641–1652. https://doi.org/10.1177/1352458518799580

Thangavelu, K. & Brunner, E. (2007). Wilcoxon–Mann–Whitney test for stratified samples and Efron's paradox dice. *J. Stat. Plann. Inference*, **137**, 720–737.

Thas, O., De Neve, J., Clement, L. & Ottoy, J.-P. (2012). Probabilistic index models. *J. R. Stat. Soc.*, **74**, 623–671.

Umlauft, M., Konietschke, F. & Pauly, M. (2017). Rank-based permutation approaches for nonparametric factorial designs. *Br. J. Math. Stat. Psychol.*, **70**, 368–390.
Umlauft, M., Placzek, M., Konietschke, F. & Pauly, M. (2019). Wild bootstrapping rank-based procedures: Multiple testing in nonparametric factorial repeated measures designs. *J. Multivar. Anal.*, **171**, 176–192.
Whitley, E. & Ball, J. (2002). Statistics review 6: Nonparametric methods. *Crit. Care*, **6**(6), 509–513.

[*Received March 2018, Revised August 2020, Accepted September 2020*]

Supporting Information

Supporting information may be found in the online version of this article.