Nonparametric Tests for the Interaction in Two-way Factorial Designs Using R

by Jos Feys

Abstract An increasing number of R packages include nonparametric tests for the interaction in two-way factorial designs. This paper briefly describes the different methods of testing and reports the resulting *p*-values of such tests on datasets for four types of designs: between, within, mixed, and pretest-posttest designs. Potential users are advised only to apply tests they are quite familiar with and not be guided by *p*-values for selecting packages and tests.

Introduction

In his book 'Discovering Statistics Using R' (Field et al., 2012), Andy Field remarked that, contrary to the popular assertion, there are robust methods that can be used to test for the interaction in mixed models. He was referring to the **WRS** package (early version of **WRS2** by Mair et al. (2015), based on Rand Wilcox's book (Wilcox, 2012)). At that time, this apparently was the only R package known to the authors for nonparametric (robust or distribution-free) tests for the interaction in factorial designs. The **nparLD** package by Noguchi et al. (2012), which offers a variety of such tests, was first published in September 2012. Since then, an increasing number of R packages have emerged with functions to run nonparametric tests for the interaction(s) in factorial designs.

The main purpose of this paper is to familiarize researchers and potential users, who have a fair knowledge of statistics, with R packages that include nonparametric tests (R functions for such tests) for the interaction in two-way factorial designs. I first shortly describe the different methods for such tests in R Packages (available at the time of writing) and then report the resulting *p*-values of the tests, applied on data of two-way between, within, and mixed factorial designs. The term *between* refers to a between-subjects independent factor (or variable), for which a different group of subjects (or units of observation) is used for each level of the factor. A *within*-subjects factor, on the other hand, is an independent factor that is manipulated by testing each participant at each level of the factor, also named *repeated measures*. *Mixed* designs are a combination of between and within factors. For the account of *p*-values, in R packages available nonparametric functions to test for the interaction were run on datasets for four types of two-way designs: 'between x between', 'within x within', 'between x within' or 'mixed', and a special case, '(between x) pretest-posttest' designs. The latter design is a common mixed design with only two levels of the within factor.

In the next section, I advise potential users not to rely on *p*-values and to justify why they chose the particular method of testing for each of the four types of designs. They should know what the chosen test does. In the concluding section the main advices are summarized and I close with the paradox Fagerland (2012) has pointed to.

Methods and R packages

The word *nonparametric* is used here in a general sense: to include all distribution-free methods that do not rely on the restrictive assumptions of parametric tests, particularly about normality of the outcome distribution and homogeneity of variances. There are some situations when it is clear that the outcome does not follow a normal distribution. These include situations when the outcome is an ordinal variable or a rank, when there are definite outliers or when the outcome has clear limits of detection. (Data with limits of detection require quite advanced special methods for analyzing (see e.g., LaFleur et al., 2011), which are not discussed here.) Tools to address assumption problems are: simulations, nonparametric tests, robust procedures, data transformation, and re-sampling. The word nonparametric is rather associated with rank tests, and 'robust' primarily refers to methods for dealing with outliers, but I use the term nonparametric for all situations.

- An account of simulation studies would, it seems to me, not fit into the purpose of the R Journal and therefore is not covered in this paper.
- Rank test and robust methods are the main topics of interest. The word *robust* can be interpreted literally. If a test is robust, the validity of the test result will not be affected by poorly structured data. Robust also has a more technical meaning. If the actual Type I error rate of a test is close to the proclaimed Type I error rate (e.g., .05) the test is considered robust.
- Data transformation is not covered in this article. Erceg-Hurn and Mirosevich (2008) remarked that transformations often fail to restore normality and homogeneity of variances, they do not

deal with outliers, they can reduce power, they sometimes rearrange the order of the means from what they were originally, and they make the interpretation of results difficult, as findings are based on the transformed rather than the original data. Data transformation should be replaced by more up-to-day methods.

- Re-sampling techniques such as *permutation* or *randomization* tests and *bootstrap* are only very concisely described here. Permutation tests use all possible distinct permutations of the dependent variable, holding the independent variables fixed. Unfortunately, a typical full permutation test is too time-consuming. An alternative is often called a randomization test. (Many authors use both terms interchangeably.) The underlying idea of randomization tests is to compare the results from the real data against the possible results if one repeatedly (e.g., 10,000 times) re-labels the data points, then see how extreme the results from the real data are, when compared against the array of alternative arrangements of the data. There are a number of R packages for randomization tests (e.g., **coin**, **lmPerm** and **perm**), but, to my knowledge, they do not readily include test for the interaction in two-way factorial designs. The `ezPerm` function from the **ez** package by Lawrence (2015) can be used for permutation tests with many types of factorial designs. (This package also has functions for visualization of the interaction using bootstrap: `ezBoot` and `ezPlot2`. Visualization methods are beyond the scope of this paper.)

  A bootstrap is a process in which data are re-sampled repeatedly (randomly with replacement and each time of the same size as the original data), and a statistic is calculated for each re-sampling to form an empirical distribution for that statistic. The **boot** package by Canty and Ripley (2016) provides extensive facilities for bootstrapping and related re-sampling methods. This package has a function for confidence intervals: `boot.ci`.

In my opinion, nonparametric tests not only have the obvious advantage of not requiring the assumption of normality or of homogeneity of variance, but also the benefit that they can be used with many different types of scales and that, when sample size is small, there may be no alternative to use a nonparametric test unless the population distribution is known exactly. Gibbons (1993) observed that ordinal scale data, notably Likert-type scales, are very common in social sciences and argued these should be analyzed with nonparametric tests.

Dealing with outliers

Rand Wilcox's book (Wilcox, 2012) and the corresponding R package **WRS2** offer robust methods for dealing with outliers: trimmed means, bootstrap (see brief description above), median tests and M-estimators.

Trimmed means This involves the calculation of the mean after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both. This number of points to be discarded is usually given as a percentage of the total number of points, but may also be given as a fixed number of points. The `t2way` and `bwtrim` functions from **WRS2** are based on 20% trimmed means, respectively for between x between and mixed (between x within) designs.

Median tests The median is a robust measure of central tendency (the mean is not), thus not influenced by outliers; therefore median tests are often chosen for dealing with outliers. The `med2way` function from **WRS2** is such a test.

M-estimators M-estimators are a general class of robust statistics which are obtained as the minima of sums of functions of the data, e.g., iterated re-weighted least-squares. As already mentioned, in the **WRS2** package, the `t2way` function computes a between x between ANOVA for trimmed means with interactions effects. The accompanying `pbad2way` performs a two-way ANOVA using M-estimators for location. With this function, the user can choose between three M-estimators for group comparisons: M-estimator of location using Huber's ψ, a modified ψ estimator, or a median. In the same package the `bwtrim` function computes a between x within (mixed) subjects ANOVA on the trimmed means. Along with this function, the `sppbi` function computes the interaction effect, using bootstrap. With this function, the user here too can choose between the same three M-estimators for group comparisons.

Ordinal data and (aligned) ranks

The vast majority of nonparametric tests are rank-based tests. Many authors have proposed their own methods of ranking to test for the interaction. A special method, the alignment of the data before ranking, was introduced early in the 1990s (see e.g., Higgins et al., 1990). Aligning implies that some estimate of a location (e.g., for the effect on a certain level of a given factor), such as the mean or median of the observations, is subtracted from each observation. These data, thus aligned according

to the desired main or interaction effect, are then ranked and parametric tests are performed on the aligned ranks. Higgins and Tashtoush (1994) offered formulas for aligning the data with completely random (between x between) designs and for repeated measures (mixed) designs.

Aligned ranks The aligned ranks tests functions `aligned.rank.transform` (from the **ART** package by Villacorta (2015)) and `art` (from the **ARTool** package by Kay and Wobbrock (2015)) can be used for between x between designs. Both functions are aligned ranks tests based on the Higgins and Tashtoush formula for completely random designs (Higgins and Tashtoush, 1994, pp. 203-204). The `art` function can also be used for within x within designs and for higher order designs. Hettmansperger also proposed a ranking method (Hettmansperger and Elmore, 2002) to test for the interaction in between x between designs which essentially corresponds to the aligned rank transform method. To my knowledge, there is no package available (yet?) implementing this method (which is quite complicated to accomplish with a simple calculator). The `npIntFactRep` function (from the **npIntFactRep** package by Feys (2015)) yields aligned ranks tests for the interaction in two-way mixed designs, based on Beasley and Zumbo (2009), and uses the Higgins and Tashtoush formula for split-plot or repeated measures designs (Higgins and Tashtoush, 1994, pp. 208) to align the data for the interaction. It lists ANOVA tables for three types of ranks: regular, Friedman, and Koch ranks.

Rank-based tests For between x between designs, the `raov` function from the **Rfit** package by Kloke and McKean (2012) is available. This package is for the rank-based analysis of linear models, a robust alternative to least squares. This `raov` test is based on reduction in dispersion for testing main effects and interaction, using an algorithm described in Hocking (1985). Gao and Alvo (2005) developed their own ranking method to test for the interaction in such designs, by comparing the sum of row ranks with the sum of column ranks. The `interaction.test` function from the **StatMethRank** package by Quinglong (2015) is an application of this method. The already mentioned **nparLD** package offers two functions for two-way designs: the `ld.f2` function for within x within and the `f1.ld.f1` function for mixed (between x within) designs. (*ld* stands for longitudinal data.) The package also offers functions for three-way designs: `f1.ld.f2` (between x within x within) and `f2.ld.f1` (between x between x within), along with functions for confidence intervals and to help researchers choose the correct function. The functions in this package are based on studies by Akritas and Brunner (see e.g., Akritas et al., 1997). The testing method defines relative treatment effects in reference to the distributions of the variables measured in the experiment. These are estimated on mean ranks. In one sense, therefore, one can think of a relative treatment effect as a generalized expectation or mean (see e.g., Shah and Madden, 2004, for an introduction to the basic concepts underlying these tests).

Resulting p-values

In this section, the resulting *p*-values are reported for various designs with concrete datasets, obtained with the appropriate R packages tests.

Between x Between

Two-way between subjects designs are dealt with first, using the 'Box-Cox' and the 'Ants-eating-lizards' data.

Box-Cox data

The **Rfit** package uses the data from Box and Cox (1964) on the survival times (10hr units) of animals in a 3 x 4 factorial experiment (n = 4 observations per cell). (The authors, Box and Cox, gave no further details about the study than that it was a biological experiment using a 3 x 4 factorial design, the factors being (a) three poisons and (b) four treatments). The distribution of the Box-Cox data is displayed in the left panel of Figure 1. The response (dependent variable) is the log survival (`logSurv`) time of the animal.

For these data, the Fligner-Killeen (median) test for the homogeneity of variances is significant (α = .05), with a p-value = .0011, as is the Shapiro-Wilk normality test, with p = .0001. (The Shapiro-Wilk test is known to be biased by sample size. With large samples, small deviations from normality yield significant results. Thus e.g., a Q–Q plot might be required for verification in addition to the test, if one really wants to address the normality issue, which is not the case here.)

As illustrated in the left panel of Figure 1, the spreading of the data in the poisons I and especially in the poisons II condition is quite larger than in the poisons III condition (which illustrates the significance of the Fligner-Killeen test), and in the B and D treatments, survival is higher than in the other two treatments.

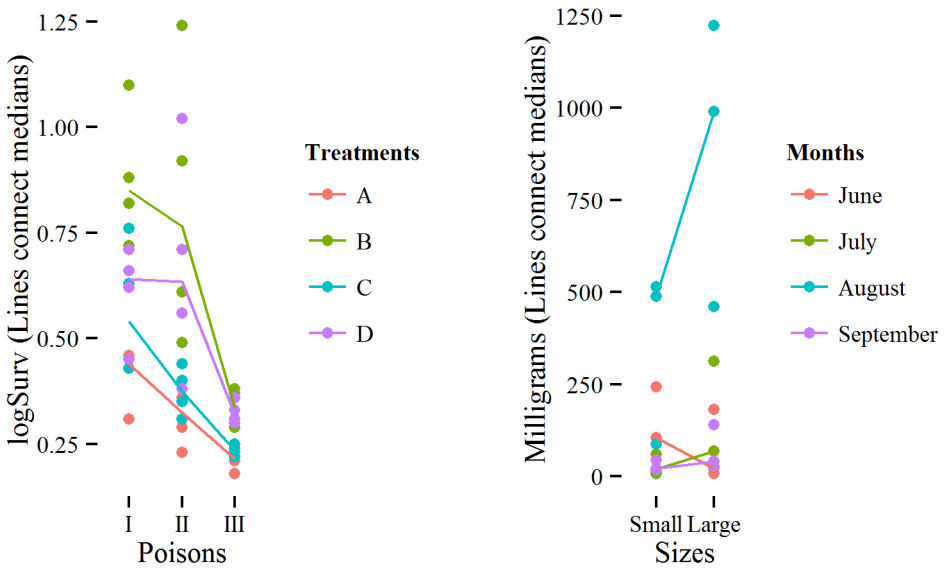

Figure 1: Distributions of the dependent variables by Between conditions for the Box-Cox (left) and Ants-eating-lizards (right) data.

In Table 1, the parametric ANOVA (*ezANOVA*, from the **ez**) on these data shows no significant interaction between treatments and poisons; neither does its permutation version *ezPerm*. The *t2way* function from **WRS2** shows no significant interaction. *pbad2way* does not run here because the covariance matrix is singular. The *med2way* function returns a significant *p*-value. The rank tests *aligned.rank.transform* (from the **ART**), *art* (from the **ARTool**) and the *raov* robust ANOVA test (from the **Rfit** package) all bring about equal and significant values, whereas the *p*-value for the ranks *interaction.test* (from the **StatMethRank**) is not significant. In the package manual, the example for the use of this function is on the Box-Cox data in matrix format.

Function	Package	Box-Cox	Ants-eating-lizards
Parametric			
ezANOVA	ez	.1123	.0617
Permutation			
ezPerm	ez	.1430	.0820
Robust			
t2way	WRS2	.0560	.3310
pbad2way	WRS2	*NA*	.5373
med2way	WRS2	.0000	.0000
Rank			
raov	Rfit	.0144	.0115
aligned.rank.transform	ART	.0168	.0688
art	ARTool	.0168	.0688
interaction.test	StatMethRank	.4913	.3959

Table 1: Resulting *p*-values of the various tests for the interaction on the Box-Cox and Ants-eating-lizards data.

Ants-eating-lizards

In the web pages material accompanying his book, in the folder on permutation tests, David Howell (Howell, 2013) illustrates the use of R scripts for permutation tests in factorial designs. He took an example from Manly (2007, p. 144). In this study, the number of ants consumed by two sizes of lizards over each of four months were observed. The distributions of milligrams of ants consumed by 24 lizards (categorized as large or small in sizes; $n = 3$ lizards per cell) during four months are displayed

in the right panel of Figure 1. It is obvious that in August, lizards – especially the large sized ones – eat most ants.

Effect	Parametric	Manly	Edgington	Still-White	ter Braak
Size	.0506	.0485	.0400	.0400	.0480
Months	.0000	.0000	.0000	.0000	.0000
S x M	.0617	.0480	.0480	.0512	.0488

Table 2: Summary of results in *p*-values from Howell (2013).

For these data, the Fligner-Killeen median test for the homogeneity of variances is not significant. Yet, the Shapiro-Wilk normality test is significant, with a *p*-value = 5.2E–06.

Howell performed several permutation tests with different approaches. The resulting *p*-values are summarized in Table 2 (as in Howell's table, *p*-values are reported only up to 4 digits after the decimal point). With the parametric test and with the Still-White approach, the *p*-values for the interaction (S x M) are 'almost' significant. (I know I should not use such terms, yet researchers typically do; this is discussed further below in the 'Choosing methods' section.) With the other approaches, the *p*-values are significant. Using the Hettmansperger method (Hettmansperger and Elmore, 2002), I calculated a very small *p*-value: 5.9E–157.

In Table 1, the resulting *p*-values for the interaction (sizes x months) with the Ants-eating-lizards data are on the right side. The parametric value using `ezANOVA` is not significant and equal to the parametric value in Table 2 (`p` = .0617). The `ezPerm` value is about the same. The `t2way` value, based on trimmed means, is not significant. The `pbad2way` function also returns a non-significant value (about .5400, depending on the ad hoc bootstrap). The `med2way` however, as for the Box-Cox data, yields a very small *p*-value. (This function only gives up to 4 digits after the decimal point.) The `raov` robust test value reveals to be significant. The aligned ranks `aligned.rank.transform` and `art` values are the same and not significant, and the *p*-value for the `interaction.test` is not significant.

Within x Within

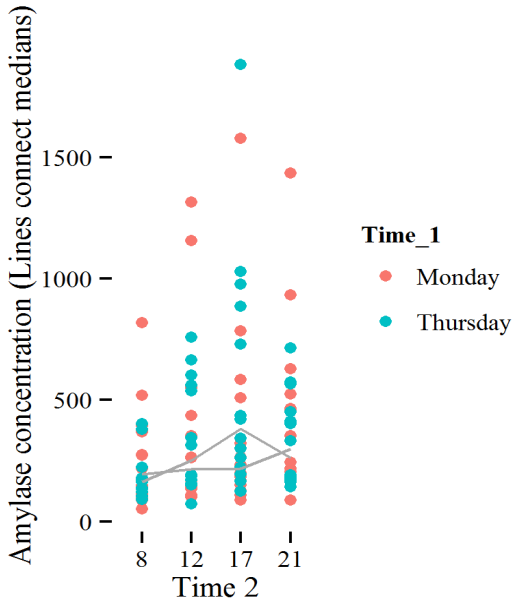

Figure 2: Distributions of the Amylase concentrations by Within conditions for the Amylase data.

Amylase data

For a two-way within (doubly repeated) subjects designs, the 'Amylase' data form the **nparLD** package were used. The data are from a longitudinal study on the concentration of α-amylase (a protein most prominent in pancreatic juice and saliva) levels (in U/ml) of the saliva from a group of 14 volunteers. Measurements were taken on 8 occasions, four times per day (8 a.m., 12 p.m., 5 p.m., 9 p.m.) and on

two days (Monday, Thursday). The distribution of the amylase concentrations in Figure 2 suggests that, on Monday, these are higher than on Thursday and that there might be an interaction between time 1 (days) and time 2 (hours). The spreading of concentrations is high at noon, and highest in the afternoon.

The Shapiro-Wilk test for normality (on the data in 'long' format) is significant with a p-value close to zero: 1.41E–12. Mauchly's sphericity test on the 8 repeated measures also reveals a significant value p = .0135. In Table 3, all p-values are significant. (The H-F value is for the Huynh-Feldt correction due

Function	Package	Amylase
Parametric		
ezANOVA	ez	.0112
– H-F corrected		.0221
Permutation		
ezPerm	ez	.0180
Rank		
art	ARTool	.0127
ld.f2	nparLD	
– Walt-type		.0025
– Anova-type		.0042

Table 3: Resulting p-values of the various tests for the interaction on the Amylase data.

to lack of sphericity.) Both values of the `ld.f2` (from **nparLD**) function are somewhat smaller than the other.

Mixed (between x within)

Three datasets were chosen for the nonparametric tests for the interaction in mixed designs: the 'Hangover', the 'Higgins', and the 'Bonate' data. The latter dataset is for a pretest-posttest mixed design, which is reviewed in the 'Pretest-Posttest' subsection.

Hangover data

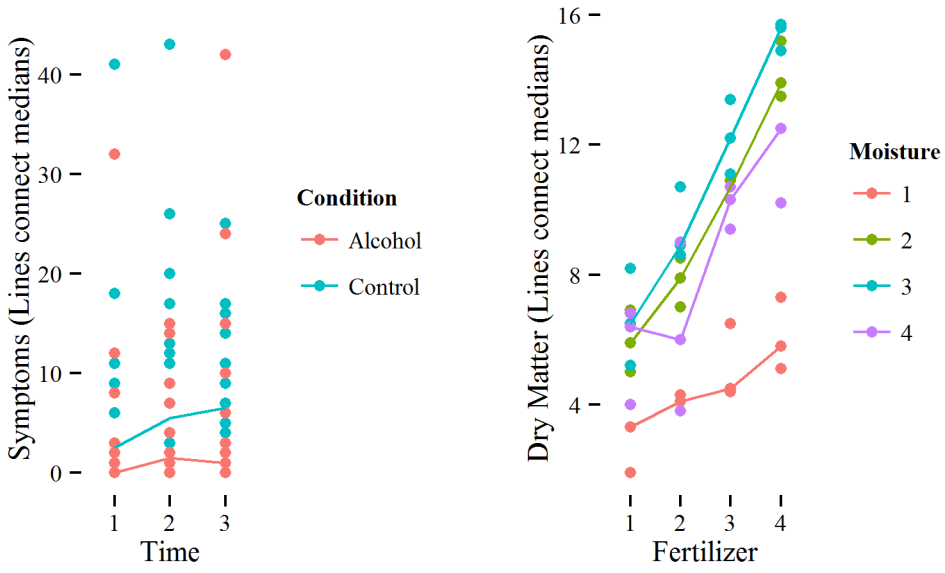

Figure 3: Distributions of the dependent variables by the Mixed conditions for the Hangover (left) and the Higgins (right) data.

The data on hangover symptoms are from Wilcox (2012, p.411). These data, also used in the **WRS2** package, come from a study on the effect of consuming alcohol, in which the number of hangover

symptoms were measured for two independent groups (N = 40, 2 x n = 20, 3 repeated measures), with each subject consuming alcohol and being measured on three different occasions. One group consisted of sons of alcoholics and the other was a control group. The distribution of this dataset is presented in the left panel of Figure 3.

The Shapiro-Wilk normality test is significant, the p-value is about zero: 4.78E–14. Mauchly's test for sphericity is not significant.

Function	Package	Hangover	Higgins
Parametric			
ezANOVA	ez	.3823	.0003
Permutation			
ezPerm	ez	.3800	.0020
Robust			
bwtrim(20%)	WRS2	.5790	*NA*
sppbi	WRS2	.8607	.0000
Rank			
npIntFactRep	npIntFactRep		
– regular		.6424	.0007
– Friedman		.7289	.0019
– Koch		.6502	.0037
art	ARTool	.3743	.0006
f1.ld.f1	nparLD		
– Walt-type		.6165	.0000
– Anova-type		.6812	.0466

Table 4: Resulting p-values of the various tests for the interaction on the Hangover and Higgins data.

None of the p-values for the interaction in the Hangover data, in the left column in Table 4, are significant.

Higgins data

The Higgins data are the Table 5 data in Higgins et al. (1990). They came from an experiment by Milliken and Johnson (1984) in which 4 peat pots, with a different (within) level of fertilizer randomly assigned to each, were placed in a tray (unit of observation). Each tray was treated with one of four different (between) moisture levels (N = 12, 4 x n = 3 trays, 4 repeated measures). The distribution of this set is displayed in the right panel of Figure 3.

Both the Shapiro-Wilk normality test and Mauchly's test for sphericity are not significant. (This implies that nonparametric tests are not really required here, but this is not an issue. Higgins et al. (1990) used these data to illustrate the aligned rank transform procedure.)

The p-values for the interaction in the Higgins data are reported in the right column of Table 4; all of them are significant. The `bwtrim` function does not run on these data because the covariance matrix is singular. `f1.ld.f1` gives the same warning; its resulting values might not be valid here.

Pretest-Posttest

This type of design is a special case of mixed designs, with only 2 within levels. Bonate (2000) thoroughly discussed the data he presented in his Table 5.4. They resulted from a study with two between groups (control and treatment; n = 10 and n = 9, respectively) and two repeated measures: pre- and posttest. In the treatment group, there was an outlier on the posttest: a value of 19 between values quite larger than 60 in the whole table. (Bonate did not give any more details about these data.)

According to Bonate, pretest-posttest data can be analyzed in several ways: (1) ANOVA on final scores alone, (2) on difference scores, (3) on percentages change scores, (4) by means of an analysis of covariance (ANCOVA) with the pre-test as covariate for the predicting group factor and the posttest as outcome variable, (5) blocking by initial scores (stratification), and (6) as repeated measures. For this design, I only review the ANCOVA, because most statisticians would agree that this should be the preferred method for analysis of pretest-posttest data (see e.g., Dimitrov and Rumrill, 2003). To test for the interaction in such a design boils down to the test for the between effect (predictor) on the posttest (criterion) after the pretest has been included in the regression model as a covariate.

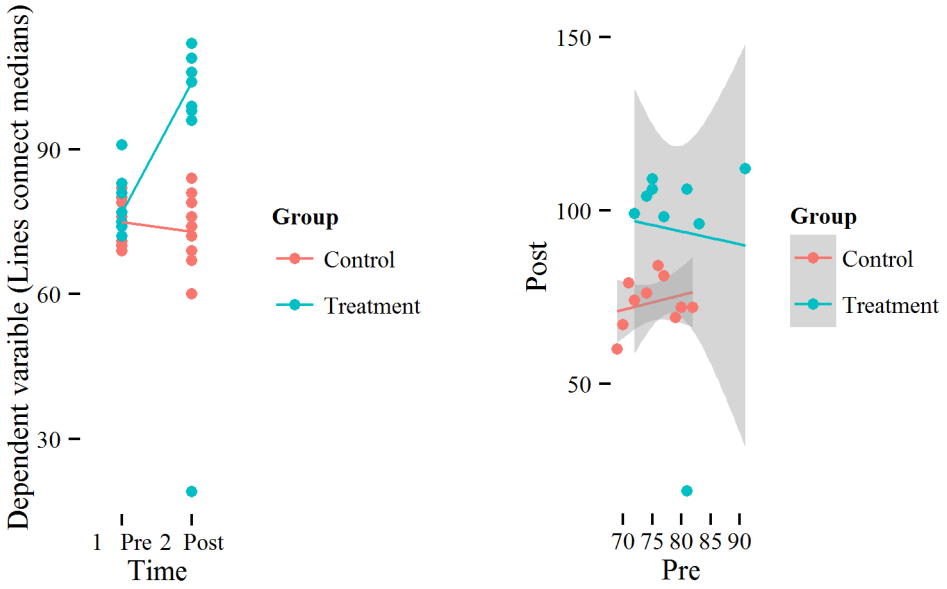

Figure 4: Distribution of the dependent variable by the Pretest-Posttest and Group conditions (left panel) and Pretest-Posttest regression plots by Group (right panel) for the Bonate data.

The Shapiro-Wilk test (on the data in 'long' format) is significant, p = .0003. Grubb's test (`grubb.test`) for one outlier (from the **outliers** package by Komsta (2011)) spots the value 19, with p = .0007.

It seems evident from Figure 4 (in the right panel) that the regression slopes are not equal between groups. (Different scales for the pre- vs. posttest were used for the plot to fit in the whole figure, despite the outlier. The shaded areas correspond to the 95% confidence intervals.) So, one of the assumptions for an ANCOVA, that of homogeneity of regression slopes, seems to be violated, most probably due to the outlier. Yet, the robust `onecova` function (from the **npsm** package by Kloke and McKean (2015), based on the their book (Kloke and McKean, 2011)), shows that the interaction group x pretest is not significant, p = .7457. Furthermore, when comparing the Pearson correlations between pre- and posttest in the two groups (r = .2683 in the control group and –.0756 in the treatment group) with the `cocor.indep.groups` function (from the **cocor** package by Diedenhofen and Musch (2015)), the resulting p-value is not significant: p = .5284. Based upon these reassuring results a nonparametric ANCOVA on these data seems justified.

Robust and rank (R)ANCOVA

Except for the outlier, in Figure 4 (left panel), all posttest values are clearly much higher in the treatment group (green dots) than in the control group (red dots). Yet, a parametric ANCOVA on these data yields a non-significant group effect (which corresponds to the interaction group x pre-posttest), with a p-value = .0576.

Bonate (2000, pp.103-106) proposed two ways for dealing with an outlier: simply removing the outlier or applying a method to minimize the influence of an observation on parameter estimations, namely the iterative re-weighted least-squares (IRWLS). He used two weight functions for the iterations: the Huber function and the bisquare. Removing the outlier from the data in this example resulted in a p-value (for the group effect) close to zero (<.0001), as did both weight functions.

Quade (1967) introduced the RANCOVA, the rank analysis of covariance. It is an ANOVA for the between effect on the residuals from the regression of the ranked posttest (criterion) on the ranked pretest (covariate). This test is quite significant here: p = .0048. A regression analysis of the ranked posttest (criterion) on ranked pretest and treatment (covariate and predictor) is another way to run a RANCOVA. It yields a comparable p-value = .0031. These values confirm Bonate's results.

The **WRS2** package offers several robust alternative for the ANCOVA: `ancov`, `ancboot`, etc. . Unfortunately, these functions fail when the number of degrees of freedom is smaller than or equal to 2, as is the case here. The `onecovahomog` function from **npsm**, which is a robust ANCOVA under homogeneous slopes, yields a p-value = .0001.

As another alternative for ANCOVA, one can run any of the many (in **WRS2** as well as in many

other packages) nonparametric or robust tests for one-way designs on variables, i.c. the residuals from the regression analysis of the posttest on the pretest. For example, the `yuen` function yields a p-value = .0001, with a trimmed mean difference between residuals of –27.24 and a 95% confidence interval: –36.10 to –18.37. The Kruskal-Wallis rank sum test returns a p-value = .0043. The Exact Wilcoxon-Mann-Whitney test yields a p-value = .0030. These tests can also be run on gain or difference scores, but many statisticians would not recommend (to say the least) analyzing such scores (see e.g., Senn, 2006).

Choosing between methods

Many statisticians (see e.g., Gardner and Altman, 1986), have warned researchers against using the accept/reject philosophy of hypothesis testing. Researchers tend to express joy on achieving a p-value of .049 and despair on finding one of 'only' .051. They seem to internalize between these values as 'right' versus 'wrong', or even of 'renewal of grants' versus 'termination of a research career'. This philosophy has led to the publication bias in journals (see e.g., Easterbrook et al., 1991), because research with statistically significant results is potentially more likely to be submitted than studies with non-significant results, with this bias of over-representation of positive results as a consequence.

The alternative, according to Gardner and Altman (1986), is to estimate the magnitude of the difference of a measured outcome between treatment groups, along with some interval that includes the population value of the difference with some specified probability: the confidence interval.

In this respect, I would advise researchers who face a choice between the increasing number of nonparametric R packages with functions for testing the interaction in factorial designs, not to compare the resulting p-values as reported in the previous section. They should certainly not go shopping for a test with the smallest p-values. Instead, they should be guided by the reason why a nonparametric test was indicated and their knowledge about how the chosen test deals with this reason why.

The above reported resulting p-values sometimes are quite different for the same datasets, depending upon the in R packages available types of tests. Since the number of available packages varies with the type of design, I discuss this choice issue for each of these separately.

Between x Between

The resulting p-values for the interaction in the between design data, summarized in Table 1, are quite divergent. With the publication strategy in mind, a researcher might very well be tempted to go for the robust median test `med2way` from **WRS2**. Yet, this test is intended to deal with outliers. If a plot of the data, e.g., a box-plot does not show any evidence of outliers, then this would not be an acceptable choice. The relevant aspect of the median test is that it only considers the position of each observation relative to the overall median, i.e., the number of times (frequency) the observation is above or below the overall median. This might not be what the researcher wants. Freidlin and Gastwirth (2000) argued the median test should be retired from general use.

For ordinal data, the researcher has a choice between the `raov`, `aligned.rank.transform`, `art` and, `interaction.test` functions; they yield comparable results. His/her choice should be inspired by the degree of knowledge he/she has or wants to invest in the rationale behind these tests. A researcher who is familiar with regression models might prefer the `raov` function; a researcher who is more acquainted with traditional ANOVA-type tests might choose one of two the aligned rank test functions. They essentially do the same thing; the `aligned.rank.transform` function has a somewhat more detailed output than `art`. The `interaction.test` yields puzzling results, because for both datasets, its p-values are quite different compared to all the other values in Table 1. Personally, I do not quite understand what this test does and therefore would not use it. Note that the permutation test `ezPerm` yields p-values not too much different from its parametric equivalent `ezANOVA`.

Within x Within

For the within x within design, all the resulting p-values in Table 3 are about equal ($p < .0200$). They all are quite significant. The two `ld.f2` type tests from **nparLD** yield the lowest values. The `art` function does not allow to specify the model for the covariance structure of the repeated measures. For doubly repeated measures designs, this might be a shortcoming. In the literature, I could not find any discussion about how to apply the aligned rank transform for doubly repeated measures. Since the `art` function uses the formula for between x between design, this might not be the right one for within x within designs. I would advise the use the `ld.f2` function here; it is better documented and was developed especially for such designs.

Mixed (between x within)

The resulting p-values in the mixed designs, displayed in Table 4, are not much dissimilar, but there are some peculiarities. For the Hangover data, the parametric test value is $p = .3823$. The nonparametric (robust and rank) p-values all are about or above .60, except for the `art` value. For the Higgins data, the anova-type value from `f1.ld.f1` seems a little odd, because it is only just under the α-level, whereas all the other values clearly are significant. As noted above, `f1.ld.f1` gave a warning; one should not choose this function here.

It is somewhat puzzling that the `art` function does not yield the same p-values as the regular aligned ranks test from **npIntFactRep**, because they supposedly both use the same procedure to align the data for the interaction, i.e. the formula for split-plot or repeated measures designs from Higgins and Tashtoush (1994, p. 208). Especially for the Hangover data the values are quite different (.6424 vs. .3743).). For the Higgins dataset, I calculated the Pearson correlation between the data aligned for the interaction with the `art` function from **ARTool** and those same data aligned for the interaction with the Higgins and Tashtoush formula for split-plot designs: this was (only) $r = .80$. When applying the Higgins and Tashtoush formula for two-way completely random designs (Higgins and Tashtoush, 1994, p. 203–204) to align the data for the interaction and then running a repeated measures ANOVA on the rank aligned data, with the *compound symmetry* options for the covariance structure (with the SAS mixed procedure described in Littell et al. (1996)), I found exactly the same value as for `art` in Table 4: $p = .3743$. This indicates that the `art` function from **ARTool** uses the wrong Higgins and Tashtoush formula for aligning the data in mixed designs. It uses the formula for between x between designs, instead of using the formula for mixed (In Higgins's terms: split-plot or repeated measures) designs. I therefore would recommend not to use the `art` function for such designs. Data with outliers should be analyzed with a robust function (or both), ordinal data can be analyzed either with `f1.ld.f1` (if it does not give a warning) or with `npIntFactRep`. Researcher's choice should be guide be his/here knowledge of these tests.

Pretest-Posttest

The parametric and permutation tests p-values for the group effect (indication the interaction) are not significant. Yet, the `onecovahomog` (robust ANCOVA) function is significant: $p = .0001$. The RANCOVAs and the nonparametric tests on residuals also all are quite significant. The p-values from the R package functions thus corroborate Bonate's results. In this context, researchers again should be guided by their knowledge about the particular tests/functions.

Conclusions

For all types of designs, the randomization (or permutation) and the parametric version of the tests yield comparable p-values. Therefore, I would not advise to use randomization tests as an genuine nonparametric alternative.

Given the sometimes quite divergent resulting p-values, potential users of nonparametric R functions to apply tests for the interaction in two-way factorials designs should be careful in their choices. They should not go shopping for the test function with the smallest p-value. Instead, a close examination and justification of the chosen function is recommended. They should know exactly what the chosen test does.

Finally, I would like to mention the paradox Fagerland (2012) has pointed to, namely that as sample sizes of research studies have increased, the use of nonparametric tests has also escalated at the expense of parametric tests. This is a paradox because parametric tests, like t-tests, are quite robust when samples sizes are large. Fagerland has shown that using nonparametric tests in large studies may provide answers to the wrong question. He stated that nonparametric tests are most useful for small-sized studies.

Bibliography

M. G. Akritas, S. F. Arnold, and E. Brunner. A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference*, 61(2):249–277, 1997. [p369]

T. M. Beasley and B. D. Zumbo. Aligned rank tests for interactions in split-plot designs: Distributional assumptions and stochastic homogeneity. *Journal of Modern Applied Statistical Methods*, 8 (1):16–50, 2009. URL http://www.soph.uab.edu/Statgenetics/People/MBeasley/Beasley-Zumbo-AlignedRanks-JMASM-2009.pdf. [p369]

P. L. Bonate. *Analysis of Pretest-Posttest Designs*. Chapman-Hall, 2000. [p373, 374]

G. Box and D. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964. URL http://fisher.osu.edu/~schroeder.9/AMIS900/Box1964.pdf. [p369]

A. Canty and B. D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2016. R package version 1.3-18. [p368]

B. Diedenhofen and J. Musch. cocor: A comprehensive solution for the statistical comparison of correlations. *PloS ONE*, 10(4):1–12, 2015. URL http://dx.doi.org/10.1371/journal.pone.0121945. [p374]

D. M. Dimitrov and P. D. Rumrill. Pretest-posttest designs and measurement of change. *Work*, 20(2): 159–165, 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12671209. [p373]

P. J. Easterbrook, J. A. Berlin, R. Gopalan, and D. R. Matthews. Publication bias in clinical research. *Lancet*, 337(8746):867–872, 1991. [p375]

D. M. Erceg-Hurn and V. M. Mirosevich. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7):591–601, 2008. [p367]

M. W. Fagerland. t-tests, non-parametric tests, and large studies – a paradox of statistical practice? *BMC Medical Research Methodology*, 12:78, 2012. URL http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-78. [p367, 376]

J. Feys. *npIntFactRep: Nonparametric Interaction Tests for Factorial Designs with Repeated Measures*, 2015. URL https://cran.r-project.org/web/packages/npIntFactRep. R package version 1.5. [p369]

A. Field, J. Miles, and Z. Field. *Discovering Statistics Using R*. SAGE, 2012. [p367]

B. Freidlin and J. L. Gastwirth. Should the median test be retired from general use? *The American Statistician*, 54(3):161–164, 2000. URL http://www.jstor.org/stable/2685584. [p375]

X. Gao and M. Alvo. A nonparametric test for interaction in two-way layout. *The Canadian Journal of Statistics*, 33(4):529–543, 2005. URL http://onlinelibrary.wiley.com/doi/10.1002/cjs.5550330405/pdf. [p369]

M. J. Gardner and D. G. Altman. Confidence intervals rather than *p*-values: Estimation rather than hypothesis testing. *British Medical Journal*, 292:746–750, 1986. [p375]

J. D. Gibbons. *Nonparametric Statistics: An Introduction*. SAGE, 1993. [p368]

T. P. Hettmansperger and R. Elmore. Tests for interaction in a two-way layout: Should they be included in a nonparametrics course? In ICOTS, editor, *Conference Proceedings*, volume 6, Cape Town, South Africa, 2002. International Association for Statistical Education. [p369, 371]

J. J. Higgins and S. Tashtoush. An aligned rank transform test for interaction. *Nonlinear World*, 1(2): 201–211, 1994. [p369, 376]

J. J. Higgins, R. C. Blair, and S. Tashtoush. The aligned rank transform procedure. In *Proceedings of the 1990 Kansas State University Conference on Applied Statistics in Agriculture*, pages 185–195, Manhattan, Kansas, 1990. Kansas State University. URL http://newprairiepress.org/agstatconference/1990/proceedings/18. [p368, 373]

R. R. Hocking. *The Analysis of Linear Models*. Brooks/Cole, 1985. [p369]

D. C. Howell. *Statistical Methods for Psychology*. Wadsworth, 8th edition, 2013. URL https://www.uvm.edu/~dhowell/. [p370, 371]

M. Kay and J. O. Wobbrock. *ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs*, 2015. URL https://cran.r-project.org/web/packages/ARTool. R package version 0.9.5. [p369]

J. D. Kloke and J. W. McKean. *Nonparametric Statistical Methods using R*. Chapman-Hall, 2011. [p374]

J. D. Kloke and J. W. McKean. Rfit: Rank-based estimation for linear models. *The R Journal*, 4(2):57–64, 2012. URL https://journal.r-project.org/archive/2012-2/RJournal_2012-2_Kloke+McKean.pdf. [p369]

J. D. Kloke and J. W. McKean. *npsm: Package for Nonparametric Statistical Methods using R*, 2015. URL https://cran.r-project.org/web/packages/npsm. R package version 0.5. [p374]

L. Komsta. *outliers: Tests for outliers*, 2011. URL https://cran.r-project.org/web/packages/outliers. R package version 0.14. [p374]

B. LaFleur, W. Lee, D. Billhiemer, C. Lockhart, J. Liu, and N. Merchant. Statistical methods for assays with limits of detection: Serum bile acid as a differentiator between patients with normal colons, adenomas, and colorectal cancer. *Journal of Carcinogenesis*, 10(12), 2011. [p367]

M. A. Lawrence. *ez: Easy Analysis and Visualization of Factorial Experiments*, 2015. URL https://cran.r-project.org/web/packages/ez. R package version 4.3. [p368]

R. C. Littell, G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. *SAS® System for Mixed Models*. SAS Institute, Cary, NC, 1996. [p376]

P. Mair, R. Wilcox, and F. Schoenbrodt. *WRS2: A Collection of Robust Statistical Methods*, 2015. URL https://cran.r-project.org/web/packages/WRS2. R package version 0.4-0. [p367]

B. F. Manly. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman-Hall, 3rd edition, 2007. [p370]

G. A. Milliken and D. E. Johnson. *Analysis of Messy Data Vol I: Designed Experiments*. Van Nostrand Reinhold Company, 1984. [p373]

K. Noguchi, M. Latif, K. Thangavelu, F. Konietschke, Y. R. Gel, and E. Brunner. *nparLD: Nonparametric Analysis of Longitudinal Data in Factorial Experiments*, 2012. URL https://cran.r-project.org/web/packages/nparLD. R package version 2.1. [p367]

D. Quade. Rank analysis of covariance. *Journal of the American Statistical Association*, 62(320):1187–1200, 1967. [p374]

L. Quinglong. *StatMethRank: Statistical Methods for Ranking Data*, 2015. URL https://cran.r-project.org/web/packages/StatMethRank. R package version 1.3. [p369]

S. Senn. Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25(24): 4334–4344, 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16921578. [p375]

D. A. Shah and L. V. Madden. Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, 94(1):33–43, 2004. URL http://apsjournals.apsnet.org/doi/pdf/10.1094/PHYTO.2004.94.1.33. [p369]

P. J. Villacorta. *ART: Aligned Rank Transform for Nonparametric Factorial Analysis*, 2015. URL https://cran.r-project.org/web/packages/ART. R package version 1.0. [p369]

R. R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier, 3rd edition, 2012. [p367, 368, 372]

Jos Feys
Faculty of Kinesiology and Rehabilitation Sciences, KU Leuven
Tervuursevest 101 - box 1500
3001 Leuven, Belgium
jos.feys@kuleuven.be