

# Robust Statistical Methods in R Using the WRS2 Package

Patrick Mair  
Harvard University

Rand Wilcox  
University of Southern California

---

## Abstract

In this manuscript we present various robust statistical methods popular in the social sciences, and show how to apply them in R using the **WRS2** package available on CRAN. We elaborate on robust location measures, and present robust  $t$ -test and ANOVA versions for independent and dependent samples, including quantile ANOVA. Furthermore, we present on running interval smoothers as used in robust ANCOVA, strategies for comparing discrete distributions, robust correlation measures and tests, and robust mediator models.

*Keywords:* robust statistics, robust location measures, robust ANOVA, robust ANCOVA, robust mediation, robust correlation.

---

## 1. Introduction

Data are rarely normal. Yet many classical approaches in inferential statistics assume normally distributed data, especially when it comes to small samples. For large samples the central limit theorem basically tells us that we do not have to worry too much. Unfortunately, things are much more complex than that, especially in the case of prominent, “dangerous” normality deviations such as skewed distributions, data with outliers, or heavy-tailed distributions.

Before elaborating on consequences of these violations within the context of statistical testing and estimation, let us look at the impact of normality deviations from a purely descriptive angle. It is trivial that the mean can be heavily affected by outliers or highly skewed distributional shapes. Computing the mean on such data would not give us the “typical” participant; it is just not a good location measure to characterize the sample. In this case one strategy is to use more robust measures such as the median or the trimmed mean and perform tests based on the corresponding sampling distribution of such robust measures.

Another strategy to deal with such violations (especially with skewed data) is to apply transformations such as the logarithm or more sophisticated Box-Cox transformations (Box and Cox 1964). For instance, in a simple  $t$ -test scenario where we want to compare two group means and the data are right-skewed, we could think of applying log-transformations within each group that would make the data “more normal”. But distributions can remain sufficiently skewed so as to result in inaccurate confidence intervals and concerns about outliers remain Wilcox (2012). Another problem with this strategy is that the respective  $t$ -test compares the log-means between the groups (i.e., the geometric means) rather than the original means.

This might not be in line anymore with the original research question and hypotheses.

Apart from such descriptive considerations, departures from normality influence the main inferential outcomes. The approximation of sampling distribution of the test statistic can be highly inaccurate, estimates might be biased, and confidence intervals can have inaccurate probability coverage. In addition, the power of classical test statistics can be relatively low.

In general, we have the following options when doing inference on small, “ugly” datasets and we are worried about basic violations. We can stay within the parametric framework and establish the sampling distribution under the null via permutation strategies. The R ([R Core Team 2016](#)) package **coin** ([Hothorn, Hornik, van de Wiel, and Zeileis 2008](#)) gives a general implementation of basic permutation strategies. However, the basic permutation framework does not provide a satisfactory techniques for comparing means ([Boik 1987](#)) or medians ([Romano 1990](#)). [Chung and Romano \(2013\)](#) summarize general theoretical concerns and limitations of permutation tests. However, they also indicate a variation of the permutation test that might have practical value.

Another option is to switch into the nonparametric testing world (see [Brunner, Domhof, and Langer 2002](#), for modern rank-based methods). Prominent examples for classical nonparametric tests taught in most introductory statistics class are the Mann-Whitney  $U$ -test ([Mann and Whitney 1947](#)), the Wilcoxon signed-rank and rank-sum test ([Wilcoxon 1945](#)), and Kruskal-Wallis ANOVA ([Kruskal and Wallis 1952](#)). However, there are well-known concerns and limitations associated with these techniques ([Wilcox 2012](#)). For example, when distributions differ, the Mann-Whitney  $U$ -test uses an incorrect estimate of the standard error.

Robust methods for statistical estimation and testing provide another good option to deal with data that are not well-behaved. Modern developments can be traced back to the 1960’s with publications by [Tukey \(1960\)](#), [Huber \(1964\)](#), and [Hampel \(1968\)](#). Measures that characterize a distribution (such as location and scale) are said to be *robust*, if slight changes in a distribution have a relatively small effect on their value ([Wilcox 2012](#), p. 23). The mathematical foundation of robust methods (dealing with quantitative, qualitative and infinitesimal robustness of parameters) makes no assumptions regarding the functional form of the probability distribution (see, e.g., [Staudte and Sheather 1990](#)). The basic trick is to view parameters as functionals; expressions for the standard error follow from the influence function. Robust inferential methods are available that perform well with relatively small sample sizes, even in situations where classic methods based on means and variances perform poorly with relatively large sample sizes. Modern robust methods have the potential of substantially increasing power even under slight departures from normality. And perhaps more importantly, they can provide a deeper, more accurate and more nuanced understanding of data compared to classic techniques based on means.

This article introduces the **WRS2** package that implements methods from the original **WRS** package ([Wilcox and Schönbrodt 2016](#)) in a more user-friendly manner. We focus on basic testing scenarios especially relevant for the social sciences and introduce these methods in an accessible way. For further technical and computational details on the original **WRS** functions as well as additional tests the reader is referred to [Wilcox \(2012\)](#).

Before we elaborate on the **WRS2** package, we give an overview of some important robust methods that are available in various R packages. In general, R is pretty well endowed with all sorts of robust regression functions and packages such as **r1m** in **MASS** ([Venables and Ripley 2002](#)), and **lmrob** and **nlrob** in **robustbase** ([Rousseeuw, Croux, Todorov, Ruckstuhl,](#)

Salibian-Barrera, Verbeke, Koller, and Maechler 2015). Robust mixed-effects models are implemented in **robustlmm** (Koller 2015) and robust generalized additive models in **robustgam** (Wong, Yao, and Lee 2014). Regarding multivariate methods, the **rrcov** package (Todorov and Filzmoser 2009) provides various implementations such as robust multivariate variance-covariance estimation and robust principal components analysis (PCA). **FRB** (Van Aelst and Willems 2013) includes bootstrap based approaches for multivariate regression, PCA and Hotelling tests, **RSKC** (Kondo 2014) functions for robust  $k$ -means clustering, and **robustDA** (Bouveyron and Girard 2015) performs robust discriminant analysis. Additional packages for robust statistics can be found on the CRAN task view on robust statistics (<https://cran.r-project.org/web/views/Robust.html>).

The article is structured as follows. After a brief introduction to robust location measures, we focus on several robust  $t$ -test/ANOVA strategies including repeated measurement designs. We then elaborate on a robust nonparametric ANCOVA involving running interval smoothers. Approaches for comparing quantiles and discrete distributions across groups are given in Section 5 before briefly elaborating on robust correlation coefficients and corresponding tests. Finally, in Section 7, we present a robust version of a mediator model. For each method presented in the article we show various applications using the respective functions in **WRS2**. The article is kept rather non-technical; for more technical details see Wilcox (2012).

## 2. Robust measures of location

A robust alternative to the arithmetic mean  $\bar{x}$  is the *trimmed mean* which discards a certain percentage at both ends of the distribution. For instance, a 20% trimmed mean cuts off 20% at the low end and 20% the high end. In R, a trimmed mean can be computed via the basic **mean** function by setting the **trim** argument accordingly. Note that if the trimming portion is set to  $\gamma = 0.5$ , the trimmed mean  $\bar{x}_t$  results in the median  $\tilde{x}$  (which by itself reflects another robust location measure).

A further robust location alternative to the mean is the *Winsorized mean*. The process of giving less weight to observations in the tails of the distribution and higher weight to the ones in the center is called *Winsorizing*. Instead of computing the mean on the original distribution we compute the mean on the Winsorized distribution. Similar to the trimmed mean, the amount of Winsorizing (i.e., the *Winsorizing level*) has to be chosen *a priori*. The **WRS2** function to compute Winsorized means is called **winmean**.

A general family of robust location measures are so called *M-estimators* (the “M” stands for “maximum likelihood-type”) which are based on a loss function to be minimized. In the simplest case we can think of a loss function of the form  $\sum_{i=1}^n (x_i - \mu)^2$ . Minimization results in a standard mean estimator  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ . Instead of such a quadratic loss we can think of a more general, differentiable distance function  $\xi(\cdot)$ :

$$\sum_{i=1}^n \xi(x_i - \mu_m) \rightarrow \min! \quad (1)$$

Let  $\Psi = \xi'(\cdot)$  denote its derivative. The minimization problem reduces to  $\sum_{i=1}^n \Psi(x_i - \mu_m) = 0$  where  $\mu_m$  denotes the *M*-estimator.

Several distance functions have been proposed in the literature. As an example, Huber (1981)

proposed the following function:

$$\Psi(x) = \begin{cases} x & \text{if } |x| \leq K \\ K \text{sign}(x) & \text{if } |x| > K \end{cases} \quad (2)$$

$K$  is the *bending constant* for which Huber proposed a value of  $K = 1.28$ . Increasing  $K$  increases sensitivity to the tails of the distribution. The estimation of  $M$ -estimators is performed iteratively (see [Wilcox 2012](#), for details) and implemented in the `mest` function.

What follows are a few examples of how to compute such simple robust location measures in R. The data vector we use is taken from [Dana \(1990\)](#) and reflects the time (in sec.) persons could keep a portion of an apparatus in contact with a specified target.

```
timevec <- c(77, 87, 88, 114, 151, 210, 219, 246, 253, 262, 296, 299, 306,
            376, 428, 515, 666, 1310, 2611)
```

Let us start with a 10% trimmed mean including standard error:

```
mean(timevec, 0.1)

## [1] 342.7059

trimse(timevec, 0.1)

## [1] 103.2686
```

Now the Winsorized mean (10% Winsorizing level) and the median with standard errors:

```
winmean(timevec, 0.1)

## [1] 380.1579

winse(timevec, 0.1)

## [1] 92.9417

median(timevec)

## [1] 262

msmedse(timevec)

## [1] 77.83901
```

As a note, `msmedse` works well when tied values never occur, but it can be highly inaccurate otherwise. Inferential methods based on a percentile bootstrap effectively deal with this issue. Finally, the Huber  $M$ -estimator with bending constant kept at its default  $K = 1.28$ .

```
mest(timevec)

## [1] 285.1576

mestse(timevec)

## [1] 52.59286
```

### 3. Robust $t$ -test and ANOVA strategies

Now we use these robust location measures in order to test for differences across groups. In the following subsections we focus on basic  $t$ -test strategies (independent and dependent groups), and various ANOVA approaches including mixed designs (i.e., between-within subjects designs).

#### 3.1. Tests on location measures for two independent groups

Yuen (1974) proposed a test statistic for a two-sample trimmed mean test which allows for unequal variances. Under the null ( $H_0: \mu_{t1} = \mu_{t2}$ ), the test statistic follows a  $t$ -distribution<sup>1</sup>. Details methods based on the median can be found in Wilcox (2012, p. 157–158). If no trimming is involved, this method reduces to Welch’s classical  $t$ -test with unequal variances (Welch 1938). Yuen’s test is implemented in the `yuen` function. There is also a bootstrap version of it (see `yuenbt`) which is suggested to be used for one-sided testing when the group sample sizes are unequal.

The example dataset consists of various soccer team statistics in five different European leagues, collected at the end of the 2008/2009 season. For the moment, let us just focus on the Spanish Primera Division (20 teams) and the German Bundesliga (18 teams). We are interested in comparing the trimmed means of goals scored per game across these two Leagues.

The group-wise boxplots and beanplots in Figure 1 visualize potential differences in the distributions. Spain has a fairly right-skewed goal distribution involving three outliers (Barcelona, Real Madrid, Atletico Madrid). In the German league, things look more balanced and symmetric. Performing a classical  $t$ -test is probably not the best option since the Spanish mean could be affected by the outliers. A safer way is to perform a two-sample test on the trimmed means. We keep the default trimming level of  $\gamma = 0.2$ .

```
yuen(GoalsGame ~ League, data = SpainGer)

## Call:
## yuen(formula = GoalsGame ~ League, data = SpainGer)
##
## Test statistic: 0.8394 (df = 16.17), p-value = 0.4135
```

<sup>1</sup>It is not suggested to use this test statistic for a  $\gamma = 0.5$  trimming level (which would result in median comparisons) since the standard errors become highly inaccurate.

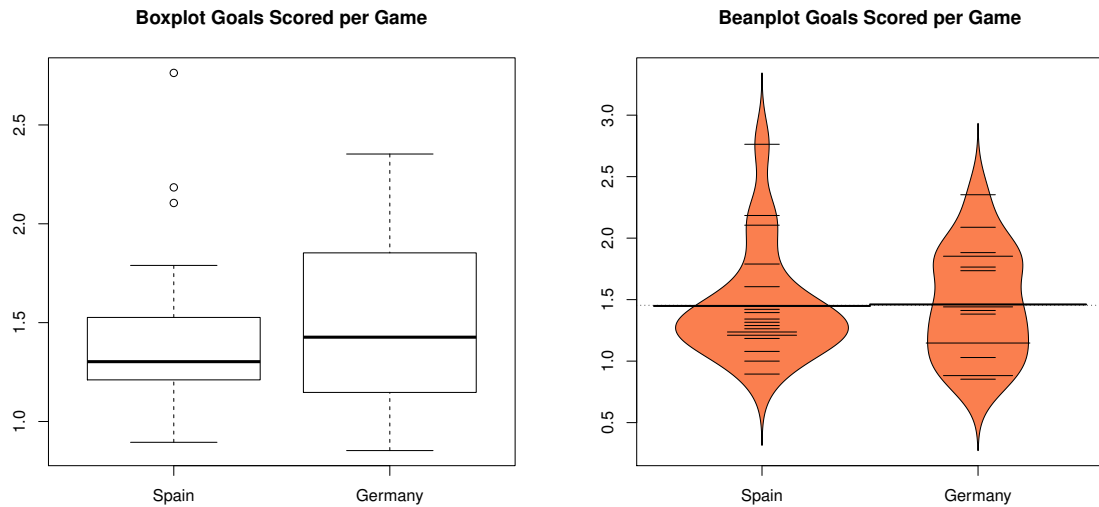


Figure 1: Left panel: boxplots for scored goals per game (Spanish vs. German league). The red dots correspond to the 20% trimmed means. Right panel: beanplots for the same setting.

```
##
## Trimmed mean difference:  -0.11494
## 95 percent confidence interval:
## -0.405      0.1751
```

The test result suggests that there are no significant differences in the trimmed means across the two leagues.

In terms of effect size, [Algina, Keselman, and Penfield \(2005\)](#) presented a robust version of Cohen's  $d$  ([Cohen 1988](#)) based on 20% trimmed means and Winsorized variances.

```
akp.effect(GoalsGame ~ League, data = SpainGer)
```

```
## [1] -0.281395
```

The same rules of thumb as for Cohen's  $d$  can be used; that is,  $|d| = 0.2$ ,  $0.5$ , and  $0.8$  correspond to small, medium, and large effects. The call above assumes equal variances across both groups. If we can not assume this, [Algina et al. \(2005\)](#) suggest to compute two effects sizes: one with the Winsorized variance of group 1 in the denominator, and another one with the Winsorized variance of group 2 in the denominator.

```
akp.effect(GoalsGame ~ League, data = SpainGer, EQVAR = FALSE)
```

```
## [1] -0.4574521 -0.2161548
```

It can happen that the two effect sizes do not lead to the same conclusions about the strength of the effect (as in our example to a certain extent). Wilcox and Tian (2011) proposed an *explanatory measure of effect size*  $\xi$  which does not suffer from this shortcoming and is generalizable to multiple groups.

```
yuen.effect.ci(GoalsGame ~ League, data = SpainGer)

## $effsize
## [1] 0.15517
##
## $CI
## [1] 0.0000000 0.6295249
```

Values of  $\hat{\xi} = 0.10, 0.30$ , and  $0.50$  correspond to small, medium, and large effect sizes. The function also gives a confidence interval for  $\hat{\xi}$ .

If we want to run a test on median differences, or more general  $M$ -estimator differences, the `pb2gen` function can be used.

```
pb2gen(GoalsGame ~ League, data = SpainGer, est = "median")

## Call:
## pb2gen(formula = GoalsGame ~ League, data = SpainGer, est = "median")
##
## Test statistic: -0.1238, p-value = 0.40067
## 95% confidence interval:
## -0.5015      0.195

pb2gen(GoalsGame ~ League, data = SpainGer, est = "onestep")

## Call:
## pb2gen(formula = GoalsGame ~ League, data = SpainGer, est = "onestep")
##
## Test statistic: -0.1181, p-value = 0.46411
## 95% confidence interval:
## -0.3786      0.1985
```

The first test is related to median differences, the second test to Huber's  $\Psi$  estimator. The results in this particular example are consistent for various robust location estimators.

### 3.2. One-way multiple group comparisons

Often it is said that  $F$ -tests are quite robust against normality violations. This is not always the case. In fact, scenarios elaborated in Games (1984), Tan (1982), Wilcox (1996) and Cressie and Whitford (1986) show that things can go wrong when applying ANOVA in situations where we have heavy-tailed distributions, unequal sample sizes, and when distributions differ in skewness. Transforming the data is not a very appealing alternative either because under general conditions this does not deal effectively with skewness issues or outliers.

The first robust ANOVA alternative presented here is a one-way comparison of multiple trimmed group means, as implemented in the `t1way` function. Let  $J$  be the number of groups. The corresponding null hypothesis is:

$$H_0 : \mu_{t1} = \mu_{t2} = \cdots = \mu_{tJ}.$$

The formula for the test statistic, which approximates an  $F$ -distribution under the null, can be found in [Wilcox \(2012, p. 293\)](#). A bootstrap version of it is provided in `t1waybt`. If no trimming is involved we end up with Welch's ANOVA version allowing for unequal variances ([Welch 1951](#)). A similar test statistic can be derived for comparing medians instead of trimmed means, implemented in the `med1way` function. When there are tied values, use instead the function `Qanova`.

Let us apply these two tests on the soccer data. This time we include all five leagues in the dataset. Figure 2 shows the corresponding boxplots and beanplots. We see that Germany and Italy have a pretty symmetric distribution, England and The Netherlands right-skewed distributions, and Spain has outliers.

In **WRS2** these robust one-way ANOVA variants can be computed as follows:

```
t1way(GoalsGame ~ League, data = eurosoccer)

## Call:
## t1way(formula = GoalsGame ~ League, data = eurosoccer)
##
## Test statistic: 1.1178
## Degrees of Freedom 1: 4
## Degrees of Freedom 2: 26.95
## p-value: 0.36875
##
## Explanatory measure of effect size: 0.22

med1way(GoalsGame ~ League, data = eurosoccer)

## Call:
## med1way(formula = GoalsGame ~ League, data = eurosoccer)
##
## Test statistic: 1.2335
## Critical value: 2.2858
## p-value: 0.24
```

None of the tests suggests a significant difference in robust goal location parameters across the leagues.

For illustration, we perform all pairwise comparisons on the same data setting. Post hoc tests on the trimmed means can be computed using the `lincon` function:



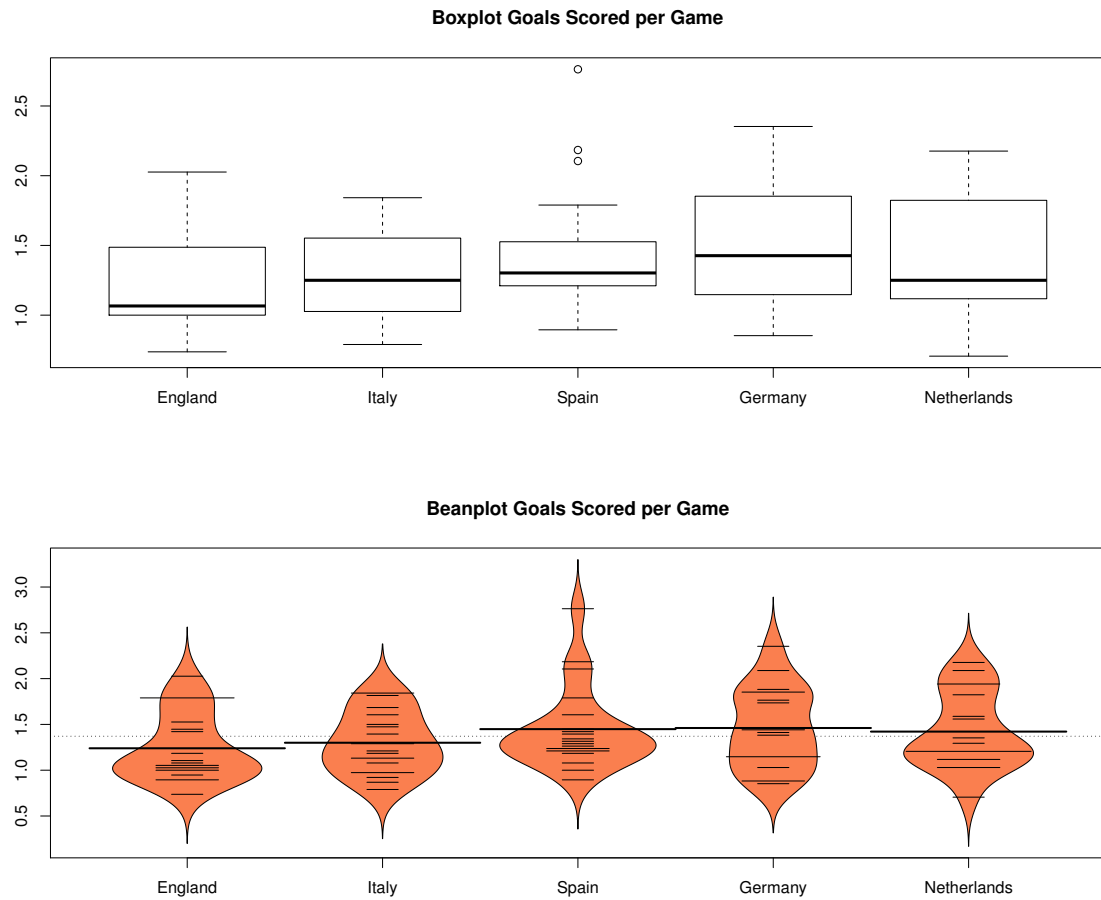


Figure 2: Top panel: Boxplots for scored goals per game across five European soccer leagues. Bottom panel: Beanplots for the same setting.

```
lincon(GoalsGame ~ League, data = eurosoccer)

## Call:
## lincon(formula = GoalsGame ~ League, data = eurosoccer)
##
##               psihat ci.lower ci.upper p.value
## England vs. Italy   -0.11184 -0.51061  0.28692 0.39635
## England vs. Spain   -0.17105 -0.50367  0.16157 0.12502
## England vs. Germany -0.28599 -0.75439  0.18241 0.07203
## England vs. Netherlands -0.22472 -0.69088  0.24145 0.14940
## Italy vs. Spain      -0.05921 -0.41380  0.29538 0.60691
## Italy vs. Germany    -0.17415 -0.65496  0.30666 0.27444
## Italy vs. Netherlands -0.11287 -0.59157  0.36583 0.47317
## Spain vs. Germany    -0.11494 -0.55124  0.32136 0.41350
```

```
## Spain vs. Netherlands   -0.05366 -0.48748  0.38015 0.69872
## Germany vs. Netherlands 0.06127 -0.47101  0.59356 0.72607
```

Post hoc tests for the bootstrap version of the trimmed mean ANOVA (**t1waybt**) are provided in **mcp2atm**.

### 3.3. Comparisons involving higher-order designs

Let us start with two-way factorial ANOVA design involving  $J$  categories for the first factor, and  $K$  categories for the second factor. The test statistic for the one-way trimmed mean comparisons can be generalized to two-way designs. The corresponding function is called **t2way**. Two-way median comparisons can be performed via **med2way** whereas for more general  $M$ -estimators, the function **pbad2way** can be applied. Note that all **WRS2** robust ANOVA functions fit the full model including all possible interactions only.

As an example we use the beer goggles dataset by [Field, Miles, and Field \(2012\)](#). This dataset is about the effects of alcohol on mate selection in night clubs. The hypothesis is that after alcohol had been consumed, subjective perceptions of physical attractiveness would become more inaccurate (*beer goggles effect*). In this dataset we have the two factors gender (24 male and 24 female students) and the amount of alcohol consumed (none, 2 pints, 4 pints). At the end of the evening the researcher took a photograph of the person the participant was chatting up. The attractiveness of the person on the photo was then evaluated by independent judges on a scale from 0-100 (response variable). Figure 3 shows the interaction plots using the median as location measure. It looks like there is some interaction going on between gender and the amount of alcohol in terms of attractiveness rating. The following code chunk computes three robust two-way ANOVA versions as well as a standard ANOVA, for comparison.

```
t2way(attractiveness ~ gender*alcohol, data = goggles)
med2way(attractiveness ~ gender*alcohol, data = goggles)
pbad2way(attractiveness ~ gender*alcohol, data = goggles, est = "onestep")
summary(aov(attractiveness ~ gender*alcohol, data = goggles))
```

For each type of ANOVA we get a significant interaction. Going back to the interaction plots in Figure 3 we see that the attractiveness of the date drops significantly for the males if they had four pints.

If we are interested in post hoc comparisons, **WRS2** provides functions for the trimmed mean version (**mcp2atm**) and the  $M$ -estimator version (**mcp2a**). Here are the results for the trimmed mean version:

```
mcp2atm(attractiveness ~ gender*alcohol, data = goggles)

## Call:
## mcp2atm(formula = attractiveness ~ gender * alcohol, data = goggles)
##
##               psihat  ci.lower  ci.upper p-value
## gender1           10.00000   -6.00223   26.00223 0.20922
```

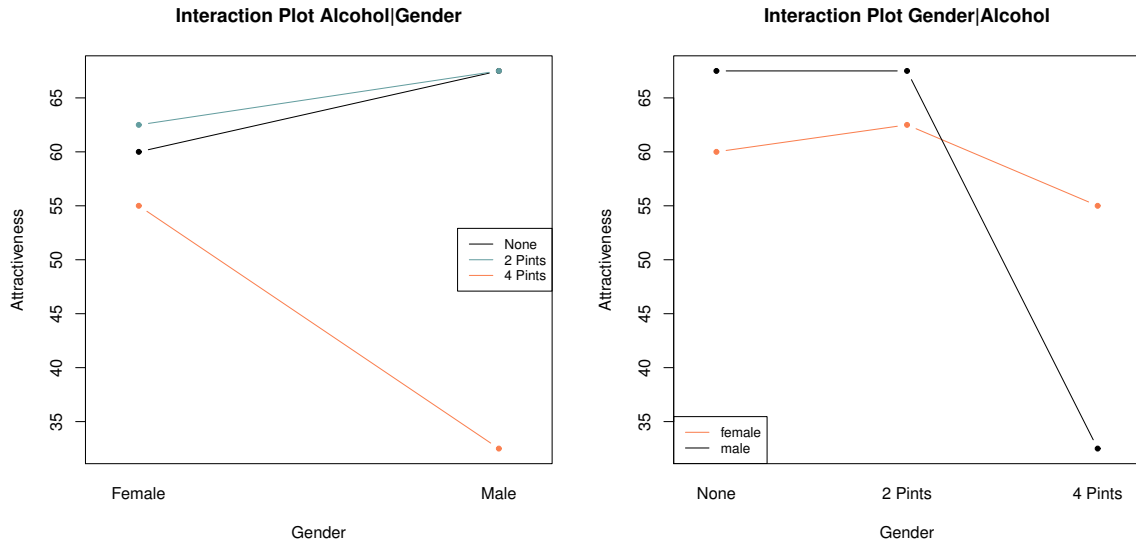


Figure 3: Interaction plot involving the median attractiveness ratings in beer goggles dataset.

```
## alcohol1      -3.33333 -20.49551  13.82885  0.61070
## alcohol2      35.83333  19.32755  52.33911  0.00003
## alcohol3      39.16667  22.46796  55.86537  0.00001
## gender1:alcohol1 -3.33333 -20.49551  13.82885  0.61070
## gender1:alcohol2 -29.16667 -45.67245 -12.66089  0.00025
## gender1:alcohol3 -25.83333 -42.53204  -9.13463  0.00080
```

The most interesting post hoc result is the `gender1:alcohol3` contrast which explains the striking 4 pint attractiveness drop for the males.

Having three-way designs, **WRS2** provides the function `t3way` for robust ANOVA based on trimmed means. The dataset we use is from [Seligman, Nolen-Hoeksema, Thornton, and Thornton \(1990\)](#). At a swimming team practice, 58 participants were asked to swim their best event as far as possible, but in each case the time reported was falsified to indicate poorer than expected performance (i.e., each swimmer was disappointed). 30 minutes later the athletes did the same performance again. The authors predicted that on the second trial more pessimistic swimmers would do worse than on their first trial, whereas optimistic swimmers would do better. The response is `ratio = Time1/Time2`. A ratio larger than 1 means that a swimmer performed better in trial 2. Figure 4 shows two separate interaction plots for male and female swimmers, using the 20% trimmed means.

Now we compute a three-way robust ANOVA on the trimmed means. For comparison, we also fit a standard three-way ANOVA (since the design is unbalanced we print out the Type II Sum-of-Squares).

```
t3way(Ratio ~ Optim*Sex*Event, data = swimming)

## Call:
```

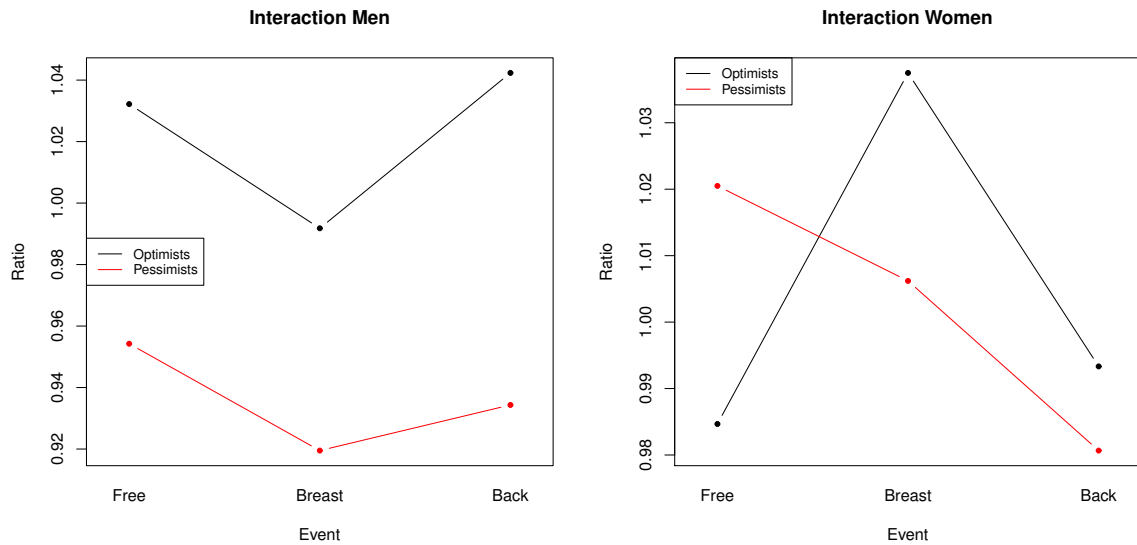


Figure 4: Interaction plot involving the trimmed means of the time ratio response for males and females separately.

```
## t3way(formula = Ratio ~ Optim * Sex * Event, data = swimming)
##
##               value p.value
## Optim          7.1799150  0.016
## Sex             2.2297985  0.160
## Event           0.3599633  0.845
## Optim:Sex       6.3298070  0.023
## Optim:Event     1.1363057  0.595
## Sex:Event       3.9105283  0.192
## Optim:Sex:Event 1.2273516  0.572

fitaov_op <- aov(Ratio ~ Optim*Sex*Event, data = swimming)
Anova(fitaov_op, type = "II")

## Anova Table (Type II tests)
##
## Response: Ratio
##               Sum Sq Df F value  Pr(>F)
## Optim          0.022923  1  6.4564 0.01449 *
## Sex             0.010084  1  2.8401 0.09871 .
## Event           0.008682  2  1.2226 0.30384
## Optim:Sex       0.018563  1  5.2283 0.02687 *
## Optim:Event     0.005076  2  0.7148 0.49464
## Sex:Event       0.010267  2  1.4459 0.24603
## Optim:Sex:Event 0.001716  2  0.2416 0.78636
```

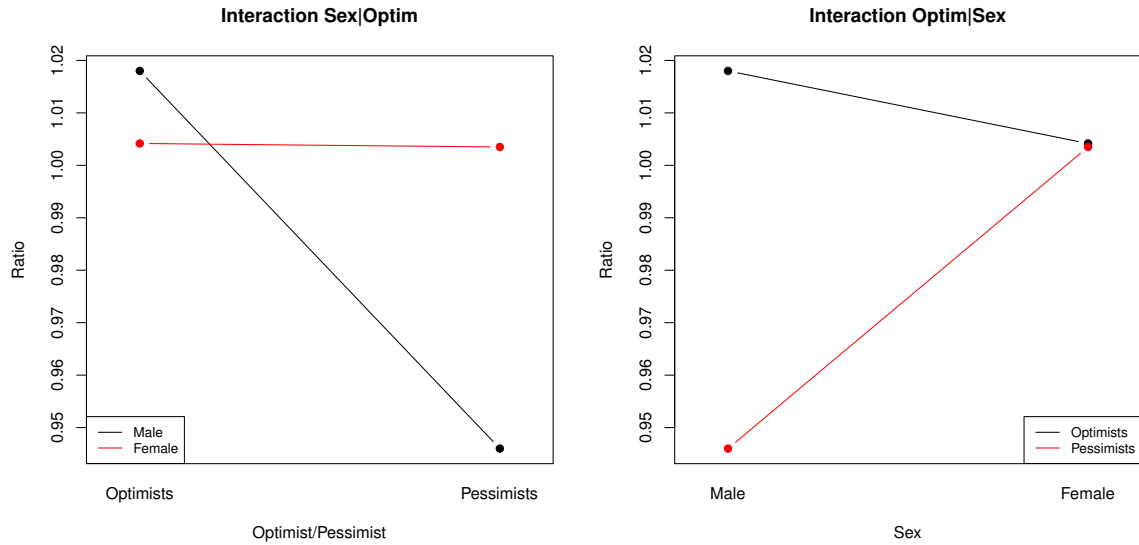


Figure 5: Interaction plot involving the trimmed means of the time ratio response for gender and optimists/pessimists (swimming style ignored).

```
## Residuals      0.163323 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The crucial effect is the **Optim:Sex** two-way interaction. Figure 5 shows the two-way interaction plot, ignoring the swimming style effect. These plots suggests that, if the swimming style is ignored, for the females it does not matter whether someone is an optimist or a pessimist. For the males, there is a significant difference in the time ratio for optimists and pessimists.

### 3.4. Repeated measurement designs

The simplest repeated measurement design is a paired samples *t*-test scenario. Yuen's trimmed mean *t*-test can be generalized to dependent data settings (i.e., within-subject designs). Details on the test statistic can be found in Wilcox (2012, p. 195–197). The corresponding R function is called **yuend** which also reports the explanatory measure of effect size.

The dataset we use for illustration is in the **MASS** package and presents data pairs involving weights of girls before and after treatment for anorexia. We use a subset of 17 girls subject to family treatment. Figure 6 presents the individual trajectories. We see that for four girls the treatment did not seem to be effective, for the remaining ones we have an increase in weight. The paired samples test on the trimmed mean differences gives a significant treatment effect which tells us that, overall, the treatment was effective (effect size can be labelled as “large”).

```
anorexiaFT <- subset(anorexia, subset = Treat == "FT")
with(anorexiaFT, yuend(Prewt, Postwt))
```

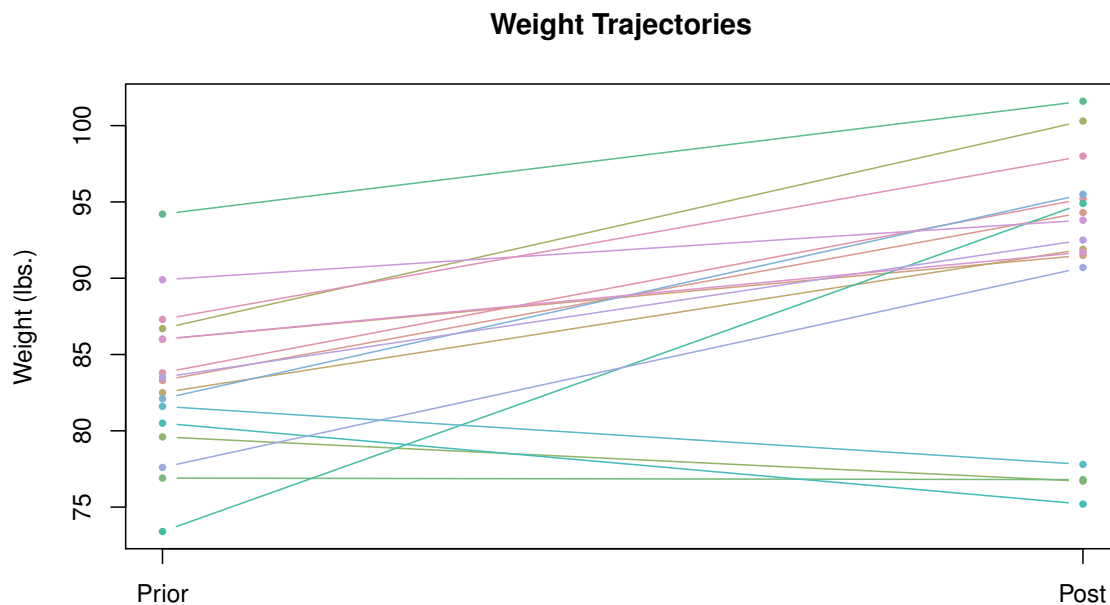


Figure 6: Individual weight trajectories of anorexic girls before and after treatment.

```
## Call:
## yuend(x = Prewt, y = Postwt)
##
## Test statistic: -3.829 (df = 10), p-value = 0.00332
##
## Trimmed mean difference: -8.56364
## 95 percent confidence interval:
## -13.5469 -3.5804
##
## Explanatory measure of effect size: 0.6
```

Let us extend this setting to more than two dependent groups. The **WRS2** package provides a robust implementation of a heteroscedastic repeated measurement ANOVA based on the trimmed means. The main function is `rmanova` with corresponding post hoc tests in `rmmcp`. The bootstrap version of `rmanova` is `rmanovab` with bootstrap post hocs in `pairdepb`. Each function for robust repeated measurement ANOVA takes three arguments; the data need to be in long format: a vector with the responses (argument: `y`), a factor for the groups (e.g., time points; argument: `groups`), and a factor for the blocks (typically a subject ID; argument: `blocks`).

The data we use to illustrate these functions is a hypothetical wine tasting dataset. There are three types of wine (A, B and C). 22 people tasted each wine five times (in a blind fold fashion). The response reflects the average ratings for each wine. Thus, each of the three

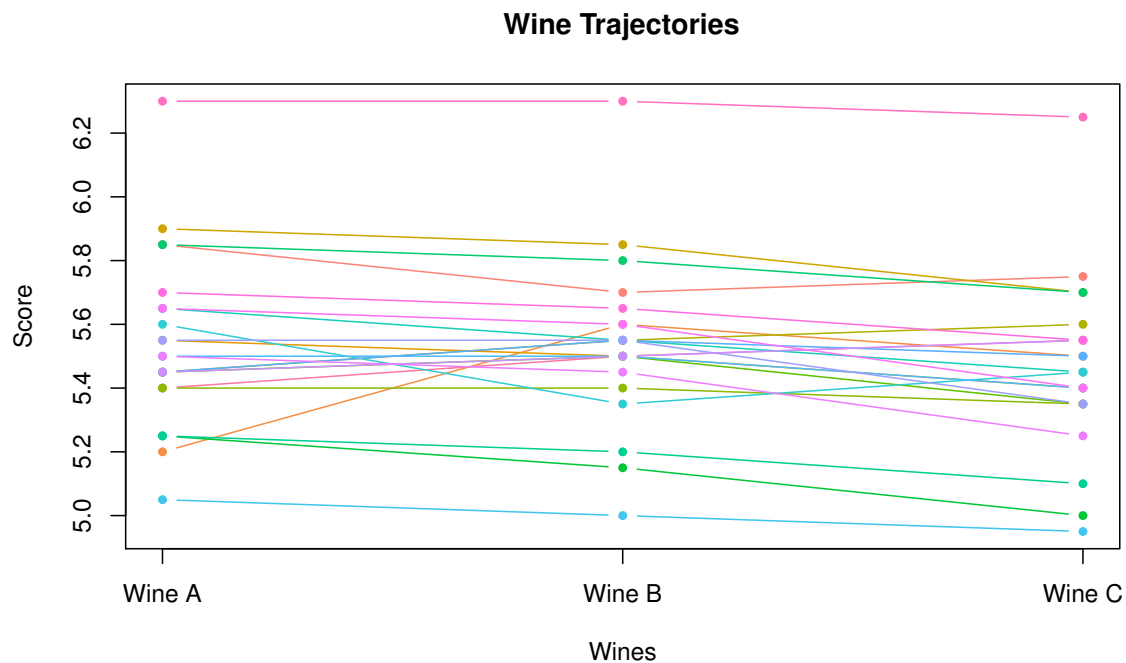


Figure 7: 22 taster trajectories for three different wines.

wines gets one score from each rater. In total, we therefore have 66 scores. The trajectories are given in Figure 7.

A robust dependent samples ANOVA on the trimmed means can be fitted as follows:

```
rmanova(y = Taste, groups = Wine, block = Taster)

## Call:
## rmanova(y = Taste, groups = Wine, blocks = Taster)
##
## Test statistic: 3.2614
## Degrees of Freedom 1: 1.61
## Degrees of Freedom 2: 20.92
## p-value: 0.06761

rmmcp(y = Taste, groups = Wine, block = Taster)

## Call:
## rmmcp(y = Taste, groups = Wine, blocks = Taster)
##
##               psihat ci.lower ci.upper p.value p.crit   sig
## Wine A vs. Wine B 0.02143 -0.02164  0.06449 0.19500 0.0500 FALSE
## Wine A vs. Wine C 0.11429  0.02148  0.20710 0.00492 0.0169  TRUE
## Wine B vs. Wine C 0.08214  0.00891  0.15538 0.00878 0.0250  TRUE
```

We see that we have a somewhat contradictory result: the global test tells us that there are no significant differences between the wines, whereas the post hoc tests suggest significant differences for the Wine C contrasts. Such results sometimes occur in small sample ANOVA applications when the global test statistic is close to the critical value.

### 3.5. Mixed designs

This subsection deals with mixed ANOVA designs, that is, we have within-subjects effects (e.g., due to repeated measurements) and between-subjects effects (group comparisons). For the parametric case, the standard `aov` function in R is able to handle such scenarios, even though in a very limited way. The `ezANOVA` function in the `ez` package (Lawrence 2015) allows for an easy specification of such models and also provides some permutation options via `ezPerm`. Since such designs belong to the mixed-effects model family, standard packages like `lme4` (Bates, Maechler, Bolker, and Walker 2015) or `nlme` (Pinheiro, Bates, DebRoy, Sarkar, and R Core Team 2015) can be applied which provide a great deal of modeling flexibility.

The main function in **WRS2** for computing a between-within subjects ANOVA on the trimmed means is `bwtrim`. For general  $M$ -estimators, the package offers the bootstrap based functions `sppba`, `sppbb`, and `sppbi` for the between-subjects effect, the within-subjects effect, and the interaction effect, respectively. Each of these functions requires the full model specification through the `formula` interface as well as an `id` argument that accounts for the within-subject structure.

The first example we use is from Wilcox (2012, p. 411). In a study on the effect of consuming alcohol, the number hangover symptoms were measured for two independent groups, 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. A representation of the dataset is given in Figure 8.

First, we fit the between-within subjects ANOVA on the 20% trimmed means:

```
bwtrim(symptoms ~ group*time, id = id, data = hangover)

## Call:
## bwtrim(formula = symptoms ~ group * time, id = id, data = hangover)
##
##           value p.value
## group       6.6087  0.0218
## time        4.4931  0.0290
## group:time  0.5663  0.5790
```

We get significant group and time effects. Second, we fit a standard between-within subjects ANOVA through `bwtrim` by setting the trimming level to 0 and check whether we get the same results as with `ezANOVA`.

```
bwtrim(symptoms ~ group*time, id = id, data = hangover, tr = 0)

## Call:
```



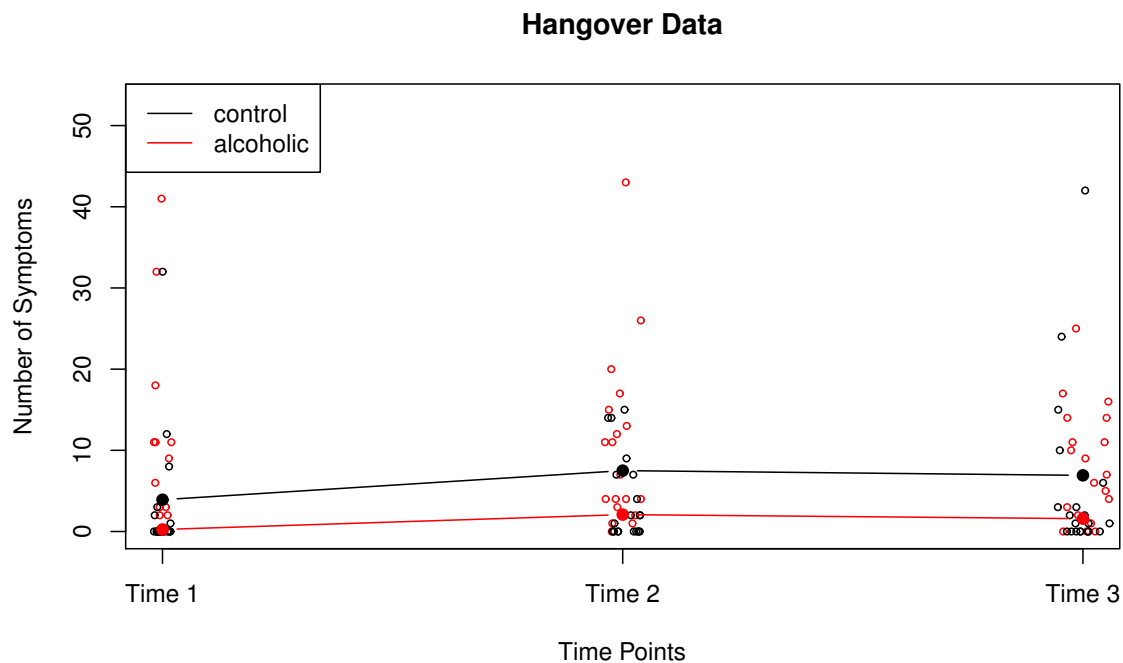


Figure 8: 20% trimmed means of the number of hangover symptoms across three time points.

```
## bwtrim(formula = symptoms ~ group * time, id = id, data = hangover,
##         tr = 0)
##
##               value p.value
## group         3.2770  0.0783
## time          0.8809  0.4250
## group:time    1.0508  0.3624
```

```
fitF <- ezANOVA(hangover, symptoms, between = group, within = time, wid = id)
fitF$ANOVA
```

##	Effect	DFn	DFd	F	p	p<.05	ges
## 2	group	1	38	3.2770015	0.07817048		0.056208518
## 3	time	2	76	0.8957333	0.41257420		0.007240111
## 4	group:time	2	76	0.9737002	0.38234407		0.007865351

Finally, we base our comparisons on Huber's  $M$ -estimator for which we have to apply three separate functions, one for each effect.

```
sppba(symptoms ~ group*time, id, data = hangover)
sppbb(symptoms ~ group*time, id, data = hangover)
```

```
sppbi(symptoms ~ group*time, id, data = hangover)
```

These tests give us a significant group effect whereas the time and interaction effects are not significant.

Due to the complexity of the hypotheses being testing using these percentile bootstrap functions, let us have a closer look using a slightly more complicated dataset. The study by McGrath (2016) looked at the effects of two forms of written corrective feedback on lexicogrammatical accuracy (**errorRatio**) in the academic writing of English as a foreign language university students. It had a  $3 \times 4$  within-by-between design with three groups (two treatment and one control; **group**) measured over four occasions (pretest/treatment, treatment, post-test, delayed post-test; **essay**).

It helps to introduce the following notations: We have  $j = 1, \dots, J$  between subjects groups (in our example  $J = 3$ ) and  $k = 1, \dots, K$  within subjects groups (e.g., time points; in our example  $K = 4$ ). Let  $Y_{ijk}$  be the response of participant  $i$  ( $i = 1, \dots, N$ ), belonging to group  $j$  on measurement occasion  $k$ .

Ignoring the group levels  $j$  for the moment,  $Y_{ijk}$  can be simplified to  $Y_{ik}$ . For two occasions  $k$  and  $k'$  we can compute the difference score  $D_{ikk'} = Y_{ik} - Y_{ik'}$ . Let  $\theta_{kk'}$  be some  $M$ -estimator associated with  $D_{ikk'}$ . In the special case of two measurement occasions (i.e.,  $K = 2$ ), we can compute a single difference. In our example with  $K = 4$  occasions we can compute  $\binom{4}{2} = 6$  such  $M$ -estimators. The null hypothesis is:

$$H_0 : \theta_{1,2} = \theta_{1,3} = \theta_{1,4} = \theta_{2,3} = \theta_{2,4} = \theta_{3,4}$$

Thus, it is tested whether the “typical” difference score (as measured by an  $M$ -estimator) between any two levels of measurement occasions is 0 (while ignoring the between-subjects groups). For the essays dataset we get:

```
sppbb(errorRatio ~ group*essay, id, data = essays)

## Call:
## sppbb(formula = errorRatio ~ group * essay, id = id, data = essays)
##
## Test statistics:
##           Estimate
## essay1-essay2 -0.083077
## essay1-essay3  0.068214
## essay1-essay4  0.003929
## essay2-essay3  0.092500
## essay2-essay4 -0.033333
## essay3-essay4 -0.065769
##
## Test whether the corresponding population parameters are the same:
## p-value: 0.362
```

The  $p$ -value suggests that we can not reject the  $H_0$  of equal difference scores.

In terms of comparisons related to the between-subjects we can think of two principles. The first one is to perform pairwise group comparisons within each  $K = 4$  measurement occasion. In our case this leads to  $4 \times \binom{3}{2}$  parameters (here, the first index relates to  $j$  and the second index to  $k$ ). We can establish the following  $K$  null hypotheses:

$$\begin{aligned} H_0^{(1)} : \quad & \theta_{1,1} = \theta_{2,1} = \theta_{3,1} \\ H_0^{(2)} : \quad & \theta_{1,2} = \theta_{2,2} = \theta_{3,2} \\ H_0^{(3)} : \quad & \theta_{1,3} = \theta_{2,3} = \theta_{3,3} \\ H_0^{(4)} : \quad & \theta_{1,4} = \theta_{2,4} = \theta_{3,4}. \end{aligned}$$

We aggregate these hypotheses into a single  $H_0$  which tests whether these  $K$  nulls are simultaneously true.

$$\begin{aligned} H_0 : \quad & \theta_{1,1} - \theta_{2,1} = \theta_{1,1} - \theta_{3,1} = \theta_{2,1} - \theta_{3,1} = \\ & \theta_{1,2} - \theta_{2,2} = \theta_{1,2} - \theta_{3,2} = \theta_{2,2} - \theta_{3,2} = \\ & \theta_{1,3} - \theta_{2,3} = \theta_{1,3} - \theta_{3,3} = \theta_{2,3} - \theta_{3,3} = \\ & \theta_{1,4} - \theta_{2,4} = \theta_{1,4} - \theta_{3,4} = \theta_{2,4} - \theta_{3,4} = 0. \end{aligned}$$

In **WRS2** this hypothesis can be tested as follows:

```
sppba(errorRatio ~ group*essay, id, data = essays, avg = FALSE)

## Call:
## sppba(formula = errorRatio ~ group * essay, id = id, data = essays,
##       avg = FALSE)
##
## Test statistics:
##
##               Estimate
## essay1 Control-Indirect  0.17664
## essay1 Control-Direct   0.10189
## essay1 Indirect-Direct -0.07475
## essay2 Control-Indirect  0.23150
## essay2 Control-Direct   0.25464
## essay2 Indirect-Direct   0.02314
## essay3 Control-Indirect  0.05614
## essay3 Control-Direct   0.18000
## essay3 Indirect-Direct   0.12386
## essay4 Control-Indirect  0.43300
## essay4 Control-Direct  -0.11489
## essay4 Indirect-Direct -0.54789
##
## Test whether the corresponding population parameters are the same:
## p-value: 0.546
```

Again, we can not reject  $H_0$ . As we see in this example, many tests have to be carried out. An alternative that seems more satisfactory in terms of type I errors is to use the average across measurement occasions, that is

$$\bar{\theta}_{j\cdot} = \frac{1}{K} \sum_{k=1}^K \theta_{jk}. \quad (3)$$

Correspondingly, in our example a null hypothesis can be formulated as

$$H_0 : \bar{\theta}_{1\cdot} = \bar{\theta}_{2\cdot} = \bar{\theta}_{3\cdot}$$

and computed via

```
sppba(errorRatio ~ group*essay, id, data = essays)

## Call:
## sppba(formula = errorRatio ~ group * essay, id = id, data = essays)
##
## Test statistics:
##           Estimate
## Control      0.2243
## Indirect     0.1054
## Direct      -0.1189
##
## Test whether the corresponding population parameters are the same:
## p-value: 0.464
```

Note that in the hangover example above we used the averaged strategy as well and since there were only two groups (**alcoholic** vs. **control**), only a single difference score was computed. Finally, let us elaborate on the **sppbi** function which performs tests on the interactions. In the **sppbb** call 6 parameters were tested and we ignored the between-subjects group structure. Now we do not further ignore the group structure and compute  $M$ -estimators based on measurement occasion differences for each group separately. In the notation below the group index is on the right hand side of the pipe symbol, the differences in measurement occasions on the left hand side. The null hypothesis is as follows:

$$\begin{aligned} H_0 : \quad & \theta_{1,2|1} - \theta_{1,3|1} = \theta_{1,4|1} - \theta_{2,3|1} = \theta_{2,4|1} - \theta_{3,4|1} = \\ & \theta_{1,2|2} - \theta_{1,3|2} = \theta_{1,4|2} - \theta_{2,3|2} = \theta_{2,4|2} - \theta_{3,4|2} = \\ & \theta_{1,2|3} - \theta_{1,3|3} = \theta_{1,4|3} - \theta_{2,3|3} = \theta_{2,4|3} - \theta_{3,4|3} = 0. \end{aligned}$$

The **WRS2** function call to test this null hypothesis is

```
sppbi(errorRatio ~ group*essay, id, data = essays)
```

```
## Call:
## sppbi(formula = errorRatio ~ group * essay, id = id, data = essays)
##
## Test statistics:
##               Estimate
## Control essay1-essay2 -0.14667
## Control essay1-essay3  0.12083
## Control essay1-essay4  0.26750
## Control essay2-essay3 -0.11778
## Control essay2-essay4 -0.02222
## Control essay3-essay4  0.09556
## Indirect essay1-essay2 -0.23600
## Indirect essay1-essay3  0.21678
## Indirect essay1-essay4  0.45278
## Indirect essay2-essay3  0.19293
## Indirect essay2-essay4 -0.07889
## Indirect essay3-essay4 -0.27182
## Direct essay1-essay2    0.10571
## Direct essay1-essay3    0.26905
## Direct essay1-essay4    0.16333
## Direct essay2-essay3   -0.20221
## Direct essay2-essay4    0.10643
## Direct essay3-essay4    0.30864
##
## Test whether the corresponding population parameters are the same:
## p-value: 0.646
```

Again, we can not reject  $H_0$ .

## 4. Robust nonparametric ANCOVA

### 4.1. Running interval smoothers

Before considering robust ANCOVA, let us elaborate on smoothers. In general, a smoother is a function that approximates the true regression line via a technique that deals with curvature in a reasonably flexible manner. Smoothing functions typically have a smoothing parameter by means of which the user can steer the degree of smoothing. If the parameter is too small, the smoothing function might overfit the data. If the parameter is too large, we might disregard important patterns. The general strategy is to find the smallest parameter so that the plot looks reasonably smooth.

A popular regression smoother is LOWESS (locally weighted scatterplot smoothing) regression which belongs to the family of nonparametric regression models and can be fitted using the `lowess` function in **stats**. The smoothers presented here involve robust location measures from Section 2 and are called *running interval smoothers*.

Let us start with the trimmed mean. We have pairs of observations  $(x_i, y_i)$ . The strategy

behind an interval smoother is to compute the  $\gamma$ -trimmed mean using all of the  $y_i$  values for which the corresponding  $x_i$ 's are close to a value of interest  $x$  (Wilcox 2012, p. 562). Let MAD be the median absolute deviation, i.e.,  $\text{MAD} = \text{median}|x_i - \tilde{x}|$ . Let  $\text{MADN} = \text{MAD}/z_{0.75}$ , where  $z_{0.75}$  represents the quantile of the standard normal distribution. The point  $x$  is said to be close to  $x_i$  if

$$|x_i - x| \leq f \times \text{MADN}.$$

Here,  $f$  as a constant which will turn out to be the smoothing parameter. As  $f$  increases, the neighborhood of  $x$  gets larger. Let

$$N(x_i) = \{j : |x_j - x_i| \leq f \times \text{MADN}\}$$

such that  $N(x_i)$  indexes all the  $x_j$  values that are close to  $x_i$ . Let  $\hat{\theta}_i$  be a robust location parameter of interest. A running interval smoother computes  $n$   $\hat{\theta}_i$  parameters based on the corresponding  $y$ -value for which  $x_j$  is close to  $x_i$ . That is, the smoother defines an interval and runs across all the  $x$ -values. Within a regression context, these estimates represent the fitted values. Eventually, we can plot the  $(x_i, \hat{\theta}_i)$  tuples into the  $(x_i, y_i)$  scatterplot which gives us the nonparametric regression fit. The smoothness of this function depends on  $f$ .

The **WRS2** package provides smoothers for trimmed means (`runmean`), general  $M$ -estimators (`rungen`), and bagging versions of general  $M$ -estimators (`runmbo`), recommended for small datasets.

Let us look at a data example, involving various  $f$  values and various robust location measures  $\hat{\theta}_i$ . We use a simple dataset from Wright and London (2009) where we are interested whether the length and heat of a chile are related. The length was measured in centimeters, the heat on a scale from 0 (“for sissys”) to 11 (“nuclear”). The left panel in Figure 9 displays smoothers involving different robust location measures. The right panel shows a trimmed mean interval smoothing with varying smoothing parameter  $f$ . We see that, at least in this dataset, there are no striking differences between the smoothers with varying location measure. The choice of the smoothing parameter  $f$  affects the function heavily, however.

## 4.2. Robust ANCOVA

ANCOVA involves a factorial design and metric covariates that were not part of the experimental manipulation. Basic ANCOVA assumes homogeneity of regression slopes across the groups when regressing the dependent variable on the covariate. A further assumption is homoscedasticity of the error terms across groups. The robust ANCOVA function in **WRS2** does not assume homoscedasticity nor homogeneity of regression slopes. In fact, it does not make any parametric assumption on the regressions at all and uses running interval smoothing (trimmed means) for each subgroup. Both nonparametric curves can be tested for subgroup differences at various points of interest along the  $x$ -continuum. This makes it very similar to what *functional data analysis* (FDA; see Ramsay and Silverman 2005) is doing. The main difference is that FDA uses smoothing splines whereas robust ANCOVA, as presented here, running interval smoothers.

The function `ancova` fits a robust ANCOVA. In its current implementation it is limited to one factor with two categories and one covariate only. A bootstrap version of it is implemented

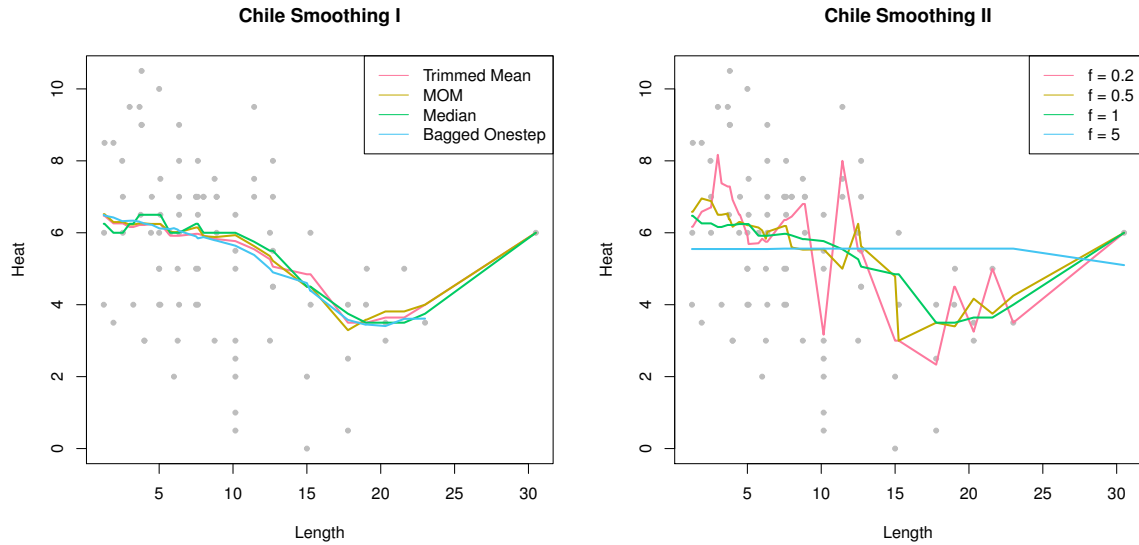


Figure 9: Left panel: smoothers with various robust location measures. Right panel: trimmed mean smoother with varying smoothing parameter  $f$ .

as well (`ancboot`). Both functions perform the running interval smoothing on the trimmed means. Yuen tests for trimmed mean differences are performed at specified design points. If the design point argument (`pts`) is not specified, the routine automatically computes five points (for details see Wilcox 2012, p. 611). It is suggested that group sizes around the design point subject to Yuen’s test should be at least 12. Regarding the multiple testing problem, the confidence intervals are adjusted to control the probability of at least one Type I error, the  $p$ -values are not.

The dataset we use to demonstrate robust ANCOVA is from Gelman and Hill (2007). It is based on data involving an educational TV show for children called “The Electric Company”. In each of four grades, the classes were randomized into treated groups and control groups. The kids in the treatment group were exposed to the TV show, those in the control group not. At the beginning and at the end of the school year, students in all the classes were given a reading test. The average test scores per class (pretest and posttest) were recorded. In this analysis we use the pretest score as covariate and are interested in possible differences between treatment and control group with respect to the posttest scores. We are interested in comparisons at six particular design points. We set the smoothing parameters to a considerably small value.

```
fitanc <- ancova(Posttest ~ Pretest + Group, fr1 = 0.3, fr2 = 0.3,
                 data = electric, pts = comppts)
fitanc

## Call:
## ancova(formula = Posttest ~ Pretest + Group, data = electric,
##        fr1 = 0.3, fr2 = 0.3, pts = comppts)
```

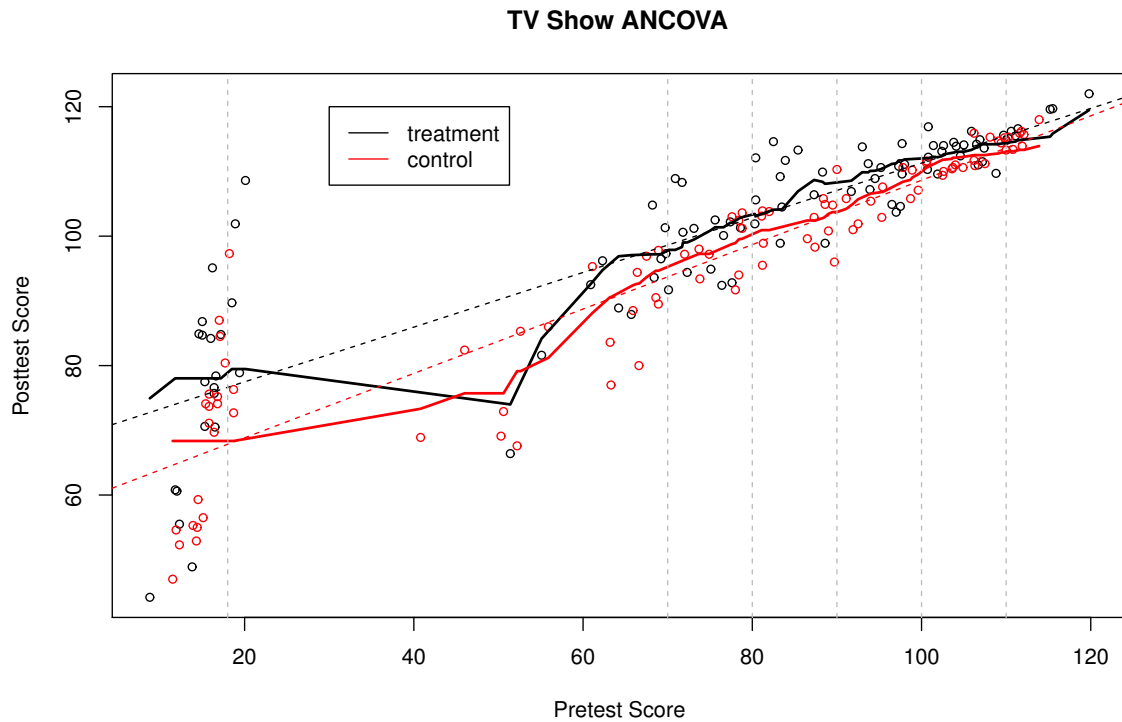


Figure 10: Robust ANCOVA fit on TV show data across treatment and control group. The nonparametric regression lines for both subgroups are shown as well as the OLS fit (dashed lines). The vertical lines show the design points our comparisons are based on.

##		n1	n2	diff	se	lower CI	upper CI	statistic	p-value
##	Pretest = 18	21	20	-11.1128	4.2694	-23.3621	1.1364	2.6029	0.0163
##	Pretest = 70	20	21	-3.2186	1.9607	-8.8236	2.3864	1.6416	0.1143
##	Pretest = 80	24	23	-2.8146	1.7505	-7.7819	2.1528	1.6079	0.1203
##	Pretest = 90	24	22	-5.0670	1.3127	-8.7722	-1.3617	3.8599	0.0006
##	Pretest = 100	28	30	-1.8444	0.9937	-4.6214	0.9325	1.8561	0.0729
##	Pretest = 110	24	22	-1.2491	0.8167	-3.5572	1.0590	1.5294	0.1380

Figure 10 shows the results of the robust ANCOVA fit. The vertical gray lines mark the design points. By taking into account the multiple testing nature of the problem, the only significant group difference we get for a pretest value of  $x = 90$ . For illustration, this plot also includes the linear regression fits for both subgroups (this is what a standard ANCOVA would do).

## 5. Other approaches for group comparisons



### 5.1. Comparing discrete distributions

Having two random variables  $X$  and  $Y$  with corresponding discrete distributions (sample space small), it might be of interest to test whether the distributions differ at each realization  $x$  and  $y$  ( $H_0: P(X = x) = P(Y = y)$ ). The function `binband` provides such an implementation allowing for both the method proposed by [Storer and Kim \(1990\)](#) and the one by [Kulinskaya, Morgenthaler, and Staudte \(2010\)](#).

Let us look at a simple artificial example. Consider a study aimed at comparing two methods for reducing shoulder pain after surgery. We provide the shoulder pain measures for each method as vector. The `binband` function compares the two distributions at each possible value in the joint sample space, here  $(1, 2, \dots, 5)$ .

```
method1 <- c(2,4,4,2,2,2,4,3,2,4,2,3,2,4,3,2,2,3,5,5,2,2)
method2 <- c(5,1,4,4,2,3,3,1,1,1,1,2,2,1,1,5,3,5)
binband(method1, method2, KMS = TRUE)
```

```
## Call:
## binband(x = method1, y = method2, KMS = TRUE)
##
## Parameter table:
```

	Value	p1.est	p2.est	p1-p2	ci.low	ci.up	p.value	p.crit
## 1	1	0.0000	0.3889	-0.3889	-0.6266	-0.1194	0.004	0.0100
## 2	2	0.5000	0.1667	0.3333	0.0201	0.6115	0.037	0.0125
## 3	3	0.1818	0.1667	0.0152	-0.2337	0.2565	0.930	0.0500
## 4	4	0.2273	0.1111	0.1162	-0.1353	0.3504	0.390	0.0167
## 5	5	0.0909	0.1667	-0.0758	-0.2969	0.1458	0.510	0.0250

Using the Kulinskaya-Morgenthaler-Staudte method (`KMS = TRUE`) we get the parameter table above and see that the distributions differ significantly at  $x, y = 1$  only. Note that the function uses Hochberg's multiple comparison adjustment to determine critical  $p$ -values.

### 5.2. Quantile comparisons

In Section 3 we described approaches for comparing robust location measures across independent groups. One such measure was the median as implemented in `pd2gen` for the two-group case and `med1way` for one-way ANOVA. Here we generalize this testing approach to arbitrary quantiles. The corresponding functions are called `qcomhd` and `Qanova` for the two-group and the multiple group case, respectively. Both of them make use of the estimator proposed by [Harrell and Davis \(1982\)](#) in conjunction with bootstrapping.

To illustrate, we use once more the soccer dataset and start comparing the German Bundesliga with the Spanish Primera Division along various quantiles.

```
fitqt <- qcomhd(GoalsGame ~ League, data = SpainGer,
               q = c(0.1, 0.25, 0.5, 0.75, 0.95), nboot = 500)
fitqt
```

```
## Call:
```

```
## qcomhd(formula = GoalsGame ~ League, data = SpainGer, q = c(0.1,
##      0.25, 0.5, 0.75, 0.95), nboot = 500)
##
## Parameter table:
##      q n1 n2  est1  est2 est1-est.2  ci.low  ci.up p.crit p.value
## 1 0.10 20 18 1.0313 0.9035      0.1278 -0.1552 0.3098 0.0100  0.268
## 2 0.25 20 18 1.1950 1.0892      0.1058 -0.1787 0.2899 0.0500  0.464
## 3 0.50 20 18 1.3109 1.4304     -0.1194 -0.5058 0.2690 0.0167  0.492
## 4 0.75 20 18 1.6220 1.8078     -0.1858 -0.6089 0.4862 0.0125  0.548
## 5 0.95 20 18 2.5160 2.2402      0.2758 -0.6043 0.8677 0.0250  0.512
```

We find no significant differences for any of the quantiles (again, the critical  $p$ -values take into account the multiple testing nature of the problem). Note that a dependent samples version of `qcomhd` is provided by the `Dqcomhd` function.

Now we extend the testing scenario above to multiple groups by considering all five leagues in the dataset and do a quartile comparison.

```
fitqa <- Qanova(GoalsGame ~ League, data = eurosoccer,
               q = c(0.25, 0.5, 0.75))
fitqa

## Call:
## Qanova(formula = GoalsGame ~ League, data = eurosoccer, q = c(0.25,
##      0.5, 0.75))
##
##           p.value  p.adj
## q = 0.25  0.0783 0.2350
## q = 0.5   0.2800 0.5600
## q = 0.75  0.7367 0.7367
```

`Qanova` adjusts the  $p$ -values using Hochberg's method (none of them significant here). For each quantile it is tested whether the test statistics are the same across the contrasts, leading to a single  $p$ -value per quantile. The contrasts itself are setup internally and the design matrix can be extracted through `fitqa$contrasts`.

## 6. Robust correlation measures

In this section we present two  $M$ -measures of correlation. The first one is the *percentage bend correlation*  $\rho_{pb}$ , a robust measure of the linear association between two random variables. When the underlying data are bivariate normal,  $\rho_{pb}$  gives essentially the same values as Pearson's  $\rho$ . In general,  $\rho_{pb}$  is more robust to slight changes in the data than  $\rho$ , similar to the robust location measures presented in Section 2. Its computation is shown in Wilcox (2012, p. 447) and involves a bending constant  $\beta$  ( $0 \leq \beta \leq 0.5$ ). It is implemented in the function `pbcor` which also performs a test on the correlation coefficient ( $H_0: \rho_{pb} = 0$ ). For illustration we use the chile dataset from Section 4.1.

```
with(chile, pbcor(length, heat))

## Call:
## pbcor(x = length, y = heat)
##
## Robust correlation coefficient: -0.3785
## Test statistic: -3.7251
## p-value: 0.00035
```

**WRS2** also provides the function `pball` for performing tests on a correlation matrix including a test statistic  $H$  which tests the global hypothesis that all percentage bend correlations in the matrix are equal to 0.

A second robust correlation measure is the *Winsorized correlation*  $\rho_w$ , which requires the specification of the amount of Winsorization. The `wincor` function can be used in a similar fashion as `pbcor`; its extension to several random variables is called `winall` and illustrated here using the hangover data from Section 3.5. We are interested in the Winsorized correlations across the three time points for the participants in the alcoholic group:

```
hangctr <- subset(hangover, subset = group == "alcoholic")
hangwide <- cast(hangctr, id ~ time, value = "symptoms")[, -1]
winall(hangwide)

## Call:
## winall(x = hangwide)
##
## Robust correlation matrix:
##      1      2      3
## 1 1.0000 0.6363 0.7049
## 2 0.6363 1.0000 0.6185
## 3 0.7049 0.6185 1.0000
##
## p-values:
##      1      2      3
## 1      NA 0.00573 0.00178
## 2 0.00573      NA 0.00750
## 3 0.00178 0.00750      NA
```

Other types of robust correlation measures are the well-known Kendall's  $\tau$  and Spearman's  $\rho$  as implemented in the basic R `cor` function.

In order to test for equality of two correlation coefficient, `twopcor` can be used for Pearson correlations and `twocor` for percentage bend or Winsorized correlations. Both functions use a bootstrap internally.

As an example, using the hangover dataset we want to test whether the time 1/time 2 correlation  $\rho_{pb1}$  of the control group is the same as the time1/time2 correlation  $\rho_{pb2}$  of the alcoholic group.

```

ct1 <- subset(hangover, subset = (group == "control" & time == 1))$symp
ct2 <- subset(hangover, subset = (group == "control" & time == 2))$symp
at1 <- subset(hangover, subset = (group == "alcoholic" & time == 1))$symp
at2 <- subset(hangover, subset = (group == "alcoholic" & time == 2))$symp
twocor(ct1, ct2, at1, at2, corfun = "pbcor", beta = 0.15)

## Call:
## twocor(x1 = ct1, y1 = ct2, x2 = at1, y2 = at2, corfun = "pbcor",
##       beta = 0.15)
##
## First correlation coefficient: 0.5628
## Second correlation coefficient: 0.5886
## Confidence interval (difference): -0.8222 0.62
## p-value: 0.96739

```

We can not reject  $H_0$ .

## 7. Robust mediation analysis

As mentioned in the Introduction, R is well-equipped with robust regression models. Here we focus on one particular approach that is especially relevant in the social sciences area: mediator models.

A simple mediator model involving a response  $Y$ , a predictor  $X$ , and a mediator  $M$  consists of the following set of regressions.

$$\begin{aligned}
 Y_i &= \beta_{01} + \beta_{11}X_i + \varepsilon_{i1}, \\
 M_i &= \beta_{02} + \beta_{12}X_i + \varepsilon_{i2}, \\
 Y_i &= \beta_{03} + \beta_{13}X_i + \beta_{23}M_i + \varepsilon_{i3}.
 \end{aligned}$$

In relation to these equations, [Baron and Kenny \(1986\)](#) laid out the following requirements for a mediating relationship:

- Significant effect of  $X$  on  $Y$  ( $\beta_{11}$ , first equation).
- Significant effect of  $X$  on  $M$  ( $\beta_{12}$ , second equation).
- The effect of  $X$  on  $Y$  when including  $M$  as well (third equation) should be reduced. If  $\beta_{13}$  is not significant anymore, we have *full mediation*, if it is still significant we have *partial mediation* and we proceed as follows.

The amount of mediation is reflected by the *indirect effect*  $\beta_{12}\beta_{23}$  (also called the *mediating effect*). Having a partial mediation situation, the state-of-the-art approach to test for mediation ( $H_0: \beta_{12}\beta_{23} = 0$ ) is to apply a bootstrap approach as proposed by [Preacher and Hayes \(2004\)](#).

In terms of a robust mediator model version, instead of OLS a robust estimation routine needs be applied to estimate the regression equations above (e.g., an  $M$ -estimator as implemented in

the `rlm` function can be used). For testing the mediating effect, [Zu and Yuan \(2010\)](#) proposed a corresponding robust approach which is implemented in **WRS2** via the `ZYmediate` function. The example we show is from [Howell \(2012\)](#) based on data by [Leerkes and Crockenberg \(2002\)](#). In this dataset ( $n = 92$ ) the relationship between how girls were raised by their own mother (**MatCare**) and their later feelings of maternal self-efficacy (**Efficacy**), that is, our belief in our ability to succeed in specific situations. The mediating variable is self-esteem (**Esteem**). All variables are scored on a continuous scale from 1 to 4.

In the first part we fit a standard mediator model with bootstrap-based testing of the mediating effect. First, we fit the three regressions as outlined above and check whether the predictor has a significant influence on the response and the mediator, respectively.

```
fit.yx <- lm(Efficacy ~ MatCare, data = Leerkes)
fit.mx <- lm(Esteem ~ MatCare, data = Leerkes)
fit.yxm <- lm(Efficacy ~ MatCare + Esteem, data = Leerkes)
fit.med <- with(Leerkes, mediation(MatCare, Esteem, Efficacy,
                                   bootstrap = TRUE, B = 500))
```

The first two regression results (not shown here) suggest that maternal care has a significant influence on the response as well as the mediator. By adding the mediator as predictor (third `lm` call), the influence of maternal care on efficacy gets lower. The Preacher-Hayes bootstrap test (we use the `mediate` function from **MBESS** ([Kelley 2016](#)) to perform the bootstrap mediation test) suggests that there is a significant mediator effect:

```
round(fit.med[1, 1:3], 4)

##           Estimate CI.Lower_Percentile CI.Upper_Percentile
##           0.0531          0.0161          0.0991
```

Now we fit the same sequence of models in a robust way. First we estimate three robust regressions using R's basic `rlm` implementation from **MASS** which uses an  $M$ -estimator. Then we perform a robust test on the mediating effect using `ZYmediate` from **WRS2**.

```
fitr.yx <- rlm(Efficacy ~ MatCare, data = Leerkes)
fitr.mx <- rlm(Esteem ~ MatCare, data = Leerkes)
fitr.yxm <- rlm(Efficacy ~ MatCare + Esteem, data = Leerkes)
with(Leerkes, ZYmediate(MatCare, Efficacy, Esteem))

## Call:
## ZYmediate(x = MatCare, y = Efficacy, med = Esteem)
##
## Mediated effect: 0.0513
## Confidence interval: 0.0159 0.0953
## p-value: 0.003
```

For the robust regression setting we get similar results as with OLS. The bootstrap based robust mediation test suggests again a significant mediator effect.

Note that robust moderator models can be fitted in a similar fashion as ordinary moderator models. Moderator models are often computed on the base of centered versions of predictor and moderator variable, including a corresponding interaction term (see, e.g., [Howell 2012](#)). In R, a classical moderator model can be fitted using `lm`. A robust version of it can be achieved by replacing the `lm` call by an `rlm` call.

## 8. Discussion

This article introduced the **WRS2** package for computing basic robust statistical methods in a user-friendly manner. Such robust models and tests are attractive when certain assumptions as required by classical statistical methods, are not fulfilled. The most important functions (with respect to social science applications) from the **WRS** package have been implemented in **WRS2**. The remaining ones are described in [Wilcox \(2012\)](#). As mentioned in the Introduction, R is already pretty well equipped with robust multivariate implementations. However, future **WRS2** updates will include robust generalizations of Hotelling's  $T$  as well as robust MANOVA.

## References

- Algina J, Keselman HJ, Penfield RD (2005). "An Alternative to Cohen's Standardized Mean Difference Effect Size: A Robust Parameter and Confidence Interval in the Two Independent Groups Case." *Psychological Methods*, **10**, 317–328.
- Baron RM, Kenny DA (1986). "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology*, **51**, 1173–1182.
- Bates D, Maechler M, Bolker BM, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.
- Boik RJ (1987). "The Fisher-Pitman Permutation Test: A Non-Robust Alternative to the Normal Theory  $F$  Test When Variances Are Heterogeneous." *British Journal of Mathematical and Statistical Psychology*, **40**, 26–42.
- Bouveyron C, Girard S (2015). **robustDA: Robust Mixture Discriminant Analysis**. R package version 1.1, URL <http://CRAN.R-project.org/package=robustDA>.
- Box GEP, Cox DR (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society B*, **26**, 211–252.
- Brunner E, Domhof S, Langer F (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. John Wiley & Sons, New York.
- Chung E, Romano JP (2013). "Exact and Asymptotically Robust Permutation Tests." *The Annals of Statistics*, **41**, 484–507.
- Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Academic Press, New York.

- Cressie NAC, Whitford HJ (1986). “How to Use the Two Sample  $t$ -Test.” *Biometrical Journal*, **28**, 131–148.
- Dana E (1990). *Salience of the Self and Salience of Standards: Attempts to Match Self to Standard*. Ph.D. thesis, Department of Psychology, University of Southern California, Los Angeles, CA.
- Field A, Miles J, Field Z (2012). *Discovering Statistics Using R*. Sage Publications, London, UK.
- Games PA (1984). “Data Transformations, Power, and Skew: A Rebuttal to Levine and Dunlap.” *Psychological Bulletin*, **95**, 345–347.
- Gelman A, Hill J (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.
- Hampel FR (1968). *Contributions to the Theory of Robust Estimation*. Ph.D. thesis, University of California, Berkeley.
- Harrell FE, Davis CE (1982). “A New Distribution-Free Quantile Estimator.” *Biometrika*, **69**, 635–640.
- Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). “Implementing a Class of Permutation Tests: The **coin** Package.” *Journal of Statistical Software*, **28**(8), 1–23. doi: [10.18637/jss.v028.i08](https://doi.org/10.18637/jss.v028.i08).
- Howell DC (2012). *Statistical Methods for Psychology*. 8th edition. Wadsworth, Belmont, CA.
- Huber PJ (1964). “Robust Estimation of Location Parameters.” *Annals of Mathematical Statistics*, **35**, 73–101.
- Huber PJ (1981). *Robust Statistics*. John Wiley & Sons, New York.
- Kelley K (2016). *MBESS: The MBESS R Package*. R package version 4.0.0, URL <https://CRAN.R-project.org/package=MBESS>.
- Koller M (2015). *robustlmm: Robust Linear Mixed Effects Models*. R package version 1.7-6, URL <http://CRAN.R-project.org/package=robustlmm>.
- Kondo Y (2014). *RSKC: Robust Sparse K-Means*. R package version 2.4.1, URL <http://CRAN.R-project.org/package=RSKC>.
- Kruskal W, Wallis WA (1952). “Use of Ranks in One-Criterion Variance Analysis.” *Journal of the American Statistical Association*, **47**, 583–621.
- Kulinskaya E, Morgenthaler S, Staudte R (2010). “Variance Stabilizing the Difference of Two Binomial Proportions.” *The American Statistician*, **64**, 350–356.
- Lawrence MA (2015). *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.3, URL <http://CRAN.R-project.org/package=ez>.
- Leerkes EM, Crockenberg SC (2002). “The Development of Maternal Self-Efficacy and Its Impact on Maternal Behavior.” *Infancy*, **3**, 227–247. doi: [10.1207/S15327078IN0302\\_7](https://doi.org/10.1207/S15327078IN0302_7).



- Mann HB, Whitney DR (1947). “On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other.” *Annals of Mathematical Statistics*, **18**, 50–60.
- McGrath D (2016). *The Effects of Comprehensive Direct and Indirect Written Corrective Feedback on Accuracy in English as a Foreign Language Students’ Writing*. Master’s thesis, Macquarie University, Sydney, Australia.
- Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-121, URL <http://CRAN.R-project.org/package=nlme>.
- Preacher KJ, Hayes AF (2004). “SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models.” *Behavior Research Methods, Instruments, and Computers*, **36**, 717–731.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Ramsay JO, Silverman BW (2005). *Functional Data Analysis*. 2nd edition. Springer Verlag, New York, NY.
- Romano JP (1990). “On the Behavior of Randomization Tests Without a Group Invariance Assumption.” *Journal of the American Statistical Association*, **85**, 686–692.
- Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Maechler M (2015). *robustbase: Basic Robust Statistics*. R package version 0.92-5, URL <http://CRAN.R-project.org/package=robustbase>.
- Seligman MEP, Nolen-Hoeksema S, Thornton N, Thornton CM (1990). “Explanatory Style as a Mechanism of Disappointing Athletic Performance.” *Psychological Science*, **1**, 143–146.
- Staudte RG, Sheather SJ (1990). *Robust Estimation and Testing*. John Wiley & Sons, New York.
- Storer BE, Kim C (1990). “Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions.” *Journal of the American Statistical Association*, **85**, 146–155.
- Tan WY (1982). “Sampling Distributions and Robustness of  $t$ ,  $F$ , and Variance-Ratio of Two Samples and ANOVA Models With Respect to Departure from Normality.” *Communications in Statistics - Theory and Methods*, **11**, 2485–2511.
- Todorov V, Filzmoser P (2009). “An Object-Oriented Framework for Robust Multivariate Analysis.” *Journal of Statistical Software*, **32**(3), 1–47. doi:10.18637/jss.v032.i03.
- Tukey JW (1960). “A Survey Sampling from Contaminated Normal Distributions.” In I Olkin, S Ghurye, W Hoeffding, W Madow, H Mann (eds.), *Contributions to Probability and Statistics*, pp. 448–503. Stanford University Press, Stanford, CA.
- Van Aelst S, Willems G (2013). “Fast and Robust Bootstrap for Multivariate Inference: The R Package **FRB**.” *Journal of Statistical Software*, **53**(3), 1–32. doi:10.18637/jss.v053.i03.
- Venables WN, Ripley BD (2002). *Modern Applied Statistics With S*. 4th edition. Springer Verlag, New York.



- Welch BL (1938). “The Significance of the Difference Between Two Means When the Population Variances are Unequal.” *Biometrika*, **29**, 350–362.
- Welch BL (1951). “On the Comparison of Several Mean Values: An Alternative Approach.” *Biometrika*, **38**, 330–336.
- Wilcox RR (1996). *Statistics for the Social Sciences*. Academic Press, San Diego, CA.
- Wilcox RR (2012). *Introduction to Robust Estimation & Hypothesis Testing*. 3rd edition. Elsevier, Amsterdam, The Netherlands.
- Wilcox RR, Schönbrodt F (2016). *A Package of R. R. Wilcox’ Robust Statistics Functions*. R package version 0.30, URL <https://github.com/nicebread/WRS/tree/master/pkg>.
- Wilcox RR, Tian T (2011). “Measuring Effect Size: A Robust Heteroscedastic Approach for Two or More Groups.” *Journal of Applied Statistics*, **38**, 1359–1368.
- Wilcoxon F (1945). “Individual Comparisons by Ranking Methods.” *Biometrics Bulletin*, **1**, 80–83.
- Wong RKW, Yao F, Lee TCM (2014). *robustgam: Robust Estimation for Generalized Additive Models*. R package version 0.1.7, URL <http://CRAN.R-project.org/package=robustgam>.
- Wright DB, London K (2009). *Modern Regression Techniques Using R*. Sage Publications, London, UK.
- Yuen KK (1974). “The Two Sample Trimmed  $t$  for Unequal Population Variances.” *Biometrika*, **61**, 165–170.
- Zu J, Yuan KH (2010). “Local Influence and Robust Procedures for Mediation Analysis.” *Multivariate Behavioral Research*, **45**, 1–44.

**Affiliation:**

Patrick Mair  
Department of Psychology  
Harvard University  
E-mail: [mair@fas.harvard.edu](mailto:mair@fas.harvard.edu)  
URL: <http://scholar.harvard.edu/mair>