Robust Statistical Methods: The R Package WRS2

Patrick Mair
Harvard University

Rand Wilcox
University of Southern California

Abstract

In this manuscript we present various robust statistical methods and show how to apply them in R using the **WRS2** package. We elaborate on robust location measures and present robust t-test versions for independent and dependent samples. We focus on robust one-way and higher order ANOVA strategies including mixed designs ("between-within subjects"). Finally, we elaborate on running interval smoothers which we use in robust ANCOVA.

Keywords: **WRS2**, robust location measures, robust ANOVA, robust ANCOVA.

1. Introduction

Data are rarely normal. Yet many classical statistical methods assume normally distributed data, especially when it comes to small samples. For large samples the central limit theorem tells us that we do not have to worry too much. Unfortunately, things are a little bit more complex than that when it comes to statistical testing, especially when we have to deal with prominent "dangerous" normality deviations such heavily skewed data, data with outliers, and heavy-tailed distribution.

Before elaborating on consequences of these violations within the context of statistcal testing and estimation, let us look at the impact of normality deviations from a purely descriptive angle. It is common knowledge that the mean can be heavily affected by outliers or highly skewed distributions. 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, once strategy is to use more robust measure such as the median or trimmed mean. Corresponding statistical tests involving such robust parameters are outside the classical statistical testing framework, however. Another strategy to deal with such violations (especially right-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 compate two group means we can think of applying log-transformations within each group which could make the data "more normal". The problem with this strategy is that a subsequent 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, violations from normality influence the results of statistical tests. The approximation of sampling distribution of the test statistic might not be proper, test results might be biased, confidence intervals not estimated in a satisfactory

manner. In addition, the power of classical test statistics becomes low. In general, we have the following options when doing inference on small, ugly datasets and we are worried about the outcomes. We can stay within the parametric framework and establish the sampling distribution via permutation strategies. The R (R Core Team 2015) package **coin** (Hothorn, Hornik, van de Wiel, and Zeileis 2008) gives a general implementation of basic permutation strategies. Another option is to perform a parametric or nonparametric bootstrap for which the **boot** package (Canty and Ripley 2015) provides a flexible framework. Alternatively, we can switch into the nonparametric testing world. Nonparametric tests have less restrictive distributional assumptions than their parametric friends. 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).

Robust methods for statistical estimation and testing provide another great option to deal with data that are not well-behaved. Historically, the first developments can be traced back to the 60'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). Robust methods are still assuming a functional form of the probability distribution but the main goal is to produce outcomes that are less sensitive to small departures from the assumed functional form. These methods are important in situations where researchers have a considerably small sample, deviating from normality. In such situations it is not a good idea to apply classical statistical tests such as t-tests, ANOVA, ANCOVA, etc. since they may deliver biased results or their power may be low.

This article introduces the **WRS2** package that implements methods from the original **WRS** package (see `https://github.com/nicebread/WRS/tree/master/pkg`) in a more user-friendly manner. We focus on basic testing scenarios especially relevant for the social sciences and introduce these methods in a simple 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, let us give an overview of some important robust methods are available in various R. packages. In general, R is pretty well endowed with all sorts of robust regression functions and packages such as `rlm` in **MASS** (Venables and Ripley 2002), `lmrob` and `nlrob` in **robustbase** (Rousseeuw, Croux, Todorov, Ruckstuhl, Salibian-Barrera, Verbeke, Koller, and Maechler 2015). The latter function performs nonlinear robust regression. 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 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 (URL: `https://cran.r-project.org/web/views/Robust.html`).

2. Robust Measures of Location

A robust alternative to the mean 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 $\tilde{x}$.

Another alternative 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 has to choosen a priori. The **WRS2** function to compute Windsorized means is `winmean`.

A general family of robust location measures are *M-estimators* (the "M" stands for "maximum likelihood-type"). The basic idea is to define a loss function to be minimized. For instance, if the loss function is $\sum_{i=1}^{n}(x_i - \mu)^2$, minimization results in the arithmetic mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Instead of such a quadratic loss we can think of a more general, differentiable distance function $\xi(\cdot)$:

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

Let $\Psi = \xi'(\cdot)$ denote its derivative. The minimization problem reduces to $\sum_{i=1}^{n} \Psi(x_i - \mu_m) = 0$ where μ_m denotes the M-estimator. Several distance functions have been proposed in the literature. As an example, Huber (see Huber 1981) proposed the following function:

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

K is the bending constant for which Huber proposed a value of $K = 1.28$. Increasing K increases efficiency when sampling from a normal distribution, but increases sensitivity to the tails of the distribution. The estimation of M-estimators is performed iteratively and implemented in the `mest` function. More details and additional distance functions can be found in Wilcox (2012).

3. Robust t-Test and ANOVA Strategies

In this section these robust location measures are used in order to test for differences across groups. 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. The test statistic is given by

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2}}, \tag{3}$$

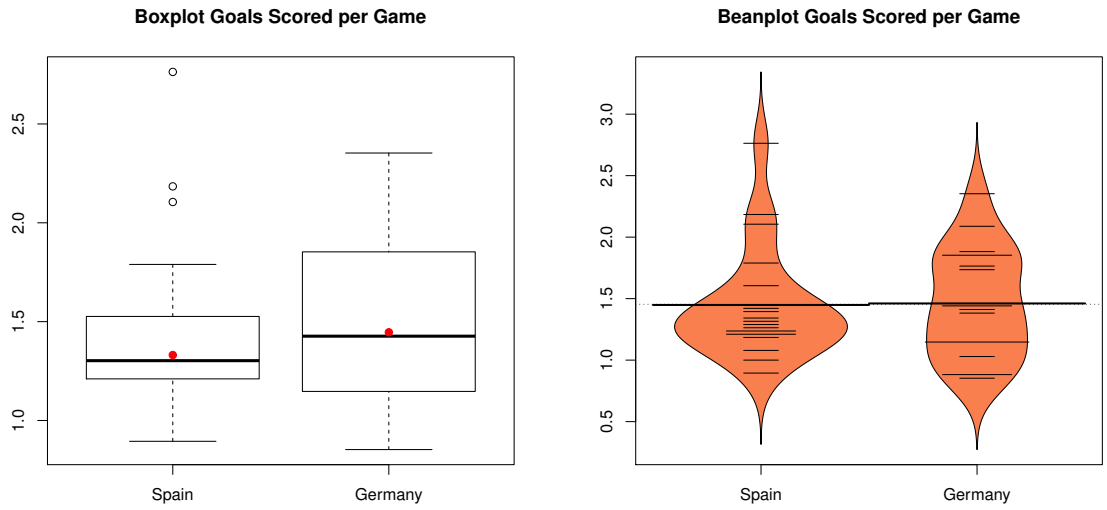

Figure 1: Left panel: boxplots for scored goals per game (Spanish vs. German league). The red dots correspond to the 20% trimmed means. Right panel: beanplots for the same setting.

which, under the null (H_0: $\mu_{t1} = \mu_{t2}$), follows a t-distribution[1]. Details on computation of the standard error and the degrees of freedom 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 in implemented in the `yuen` function. There is also a bootstrap version of it (see `yuenbt`) which is suggested to use for one-sided testing when the group sample sizes are unequal.

Let us look at an example. The dataset comprises 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 t-test on the group means could be risky, since the Spanish mean could be affected by the outliers. A saver way is to perform a test on the trimmed means. We keep the default trimming level of $\gamma = 0.2$.

Running a two-sample trimmed mean test suggests that there are no significant differences in the trimmed means across the two leagues:

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

## Call:
```

[1]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.

```
## yuen(formula = GoalsGame ~ League, data = SpainGer)
##
## Test statistic: 0.8394 (df = 16.17), p-value = 0.4135
##
## Trimmed mean difference:  -0.11494
## 95 percent confidence interval:
## -0.405     0.1751
```

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 = 0.44073
## 95 percent confidence interval:
## -0.5132     0.1749

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

## Call:
## pb2gen(formula = GoalsGame ~ League, data = SpainGer, est = "onestep")
##
## Test statistic: -0.1181, p-value = 0.44407
## 95 percent confidence interval:
## -0.354     0.2146
```

The first test related to median differences, the second test to Huber's Ψ 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 violations. This is not always the case. In fact, discussions and examples given 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 since, as in a t-test setting, we end up comparing geometric means.

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

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

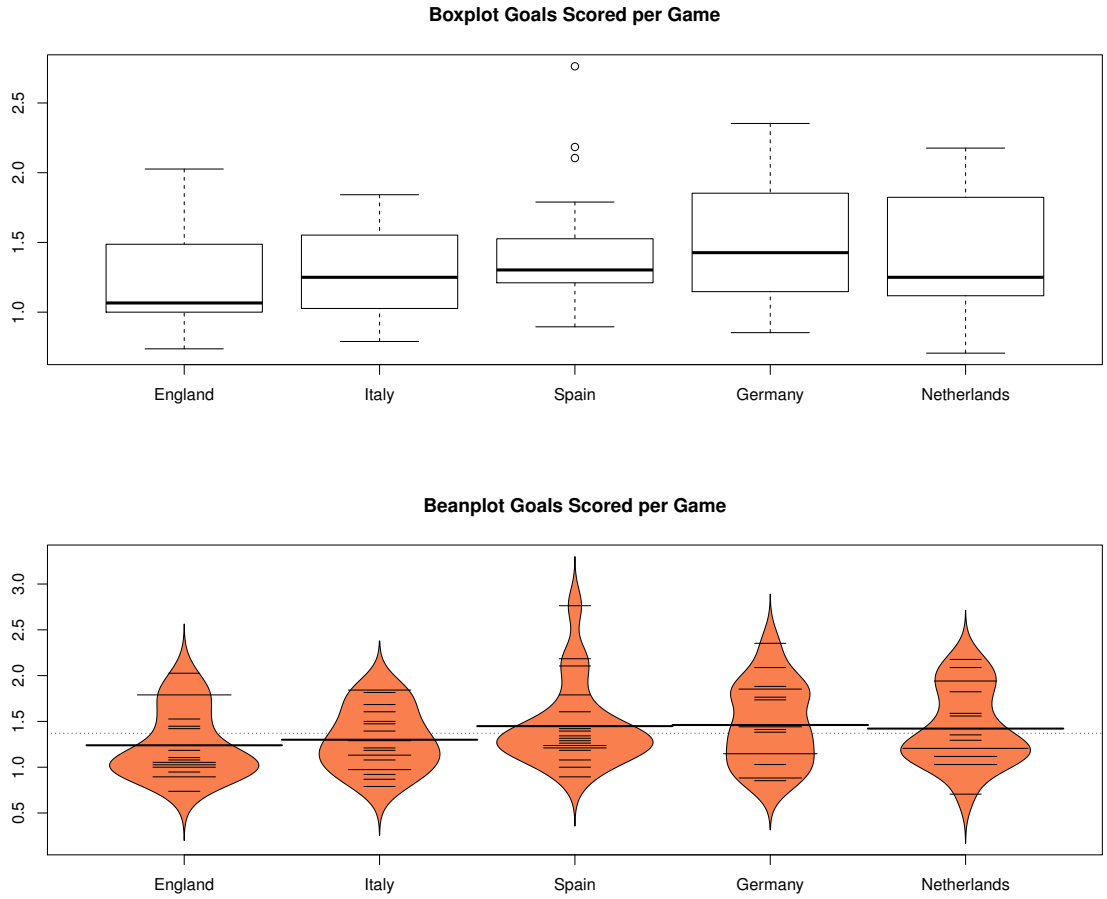

Figure 2: Left panel: Boxplots for scored goals per game (Spanish vs. German league). Right panel: Beanplots for the same setting.

The corresponding test statistic which approximates an F-distribution under the null, is quite complicated and can be found in Wilcox (2012, p. 293). A bootstrap version 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. Let us apply these two tests on the soccer dataset. This time we include all five leagues. Figure 2 shows the corresponding boxplots and beanplots. We see that Germany and Italy have a pretty symmetric distribution, England and The Nethderlands 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

med1way(GoalsGame ~ League, data = eurosoccer)

## Call:
## med1way(formula = GoalsGame ~ League, data = eurosoccer)
##
## Test statistic: 1.2335
## Critical value: 2.2315
## p-value: 0.252
```

Again, none of the tests suggests a significant difference in robust location parameters across groups. For illustration, let us just perform all pairwise comparisons on the same data setting. Post hoc tests on the trimmed means can be computed using the `lincon` function:

```
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 `mcppb20`.

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 easily generalized to two-way designs. The corresponding function is called `t2way`. Median comparisons can be performed via `med2way` whereas for more general

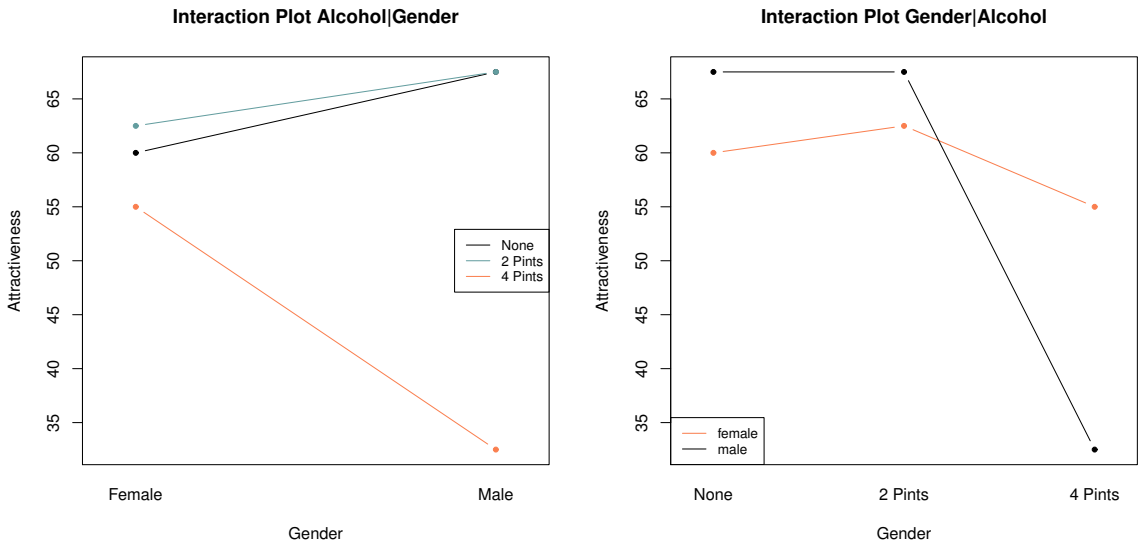

Figure 3: Interaction plot involving the median attractiveness ratings in beer goggles dataset.

M-estimators, the function `pbad2way` does the job. Note that all **WRS2** robust ANOVA functions fit the full model including all possible interactions only.

As an example we use the infamous 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 femals 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)

## Call:
## t2way(formula = attractiveness ~ gender * alcohol, data = goggles)
##
##                  value p.value
## gender          1.6667   0.209
## alcohol        48.2845   0.001
## gender:alcohol 26.2572   0.001

med2way(attractiveness ~ gender*alcohol, data = goggles)
```

```
## Call:
## med2way(formula = attractiveness ~ gender * alcohol, data = goggles)
##
##                  value p.value
## gender          6.8444  0.0089
## alcohol         4.8207  0.0081
## gender:alcohol 12.9593  0.0015

pbad2way(attractiveness ~ gender*alcohol, data = goggles, est = "onestep")

## Call:
## pbad2way(formula = attractiveness ~ gender * alcohol, data = goggles,
##     est = "onestep")
##
##                p.value
## gender           0.177
## alcohol          0.000
## gender:alcohol   0.000

summary(aov(attractiveness ~ gender*alcohol, data = goggles))

##                Df Sum Sq Mean Sq F value   Pr(>F)
## gender          1    169   168.7   2.032    0.161
## alcohol         2   3332  1666.1  20.065 7.65e-07 ***
## gender:alcohol  2   1978   989.1  11.911 7.99e-05 ***
## Residuals      42   3488    83.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In each case 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 we give 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
## 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
```

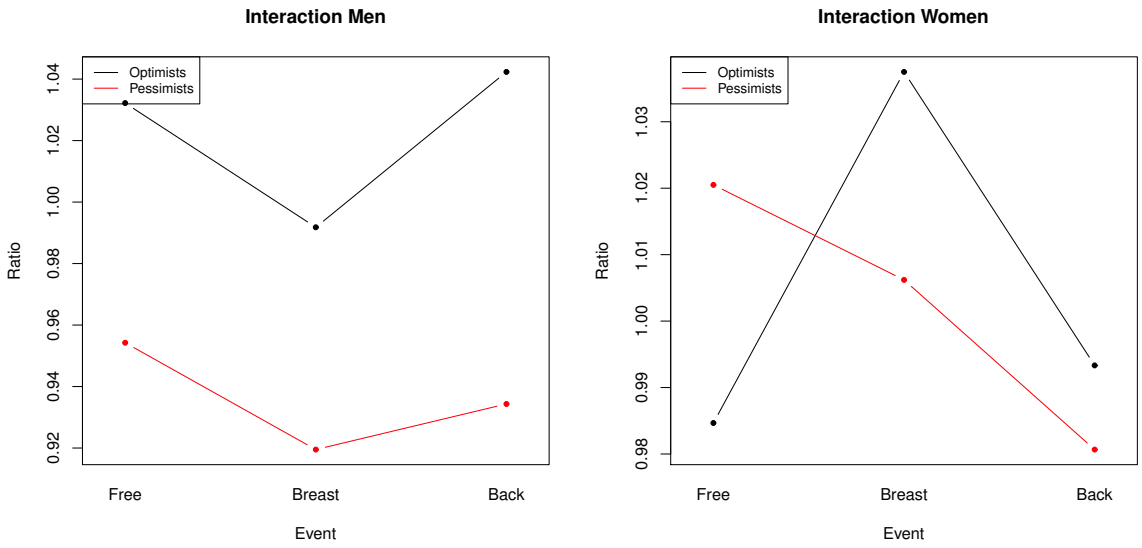

Figure 4: Interaction plot involving the trimmed means of the time ratio response for males and females separately.

```
## 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.

Now we move on to higher-order designs. **WRS2** provides the function `t3way` for robust three-way 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 that was reported was falsified to indicate poorer than expected performance (i.e., each swimmer was disappointed). 30 min later, they did the same performance. 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, involving 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:
## 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
## 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 in Equation (3) can be generalized to

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2 - 2d_{12}}}. \tag{4}$$

Expressions for the standard deviations can be found in Wilcox (2012, p. 196). The corresponding R function is called `yuend`. 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

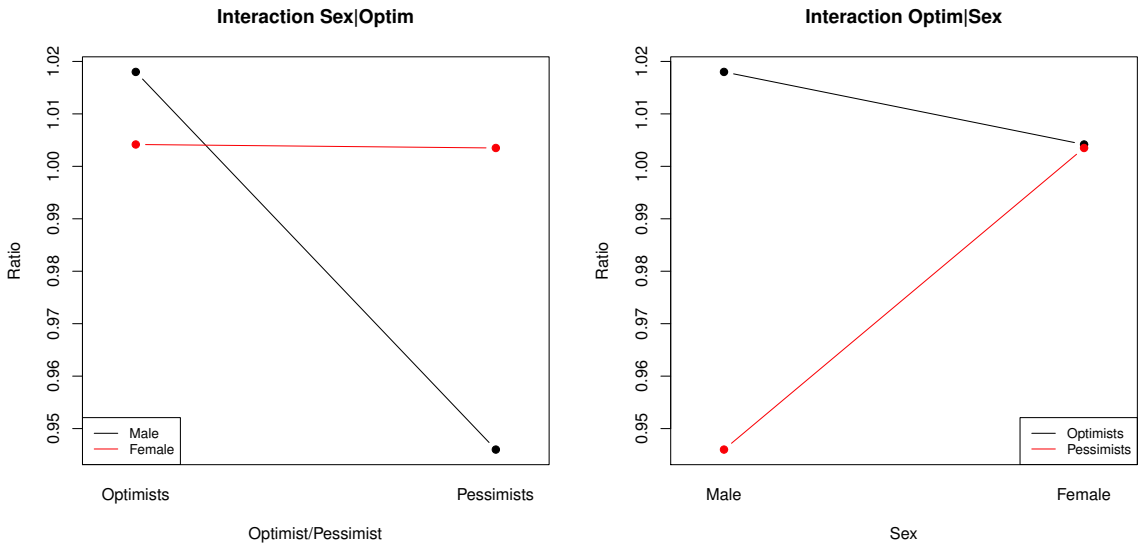

Figure 5: Interaction plot involving the trimmed means of the time ratio response for gender and optimists/pessimists (swimming style ignored).

samples test on the trimmed mean differences gives a significant treatment effect which tells us that, overall, the treatment was effective.

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

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

Let us extend this setting to more than two dependent categories. 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 the functions is a hypothetical wine tasting dataset. There are three types of wine (A, B and C). 22 people tasted each of the three wines (in a blind fold fashion), five times each. The response reflects the average ratings for each

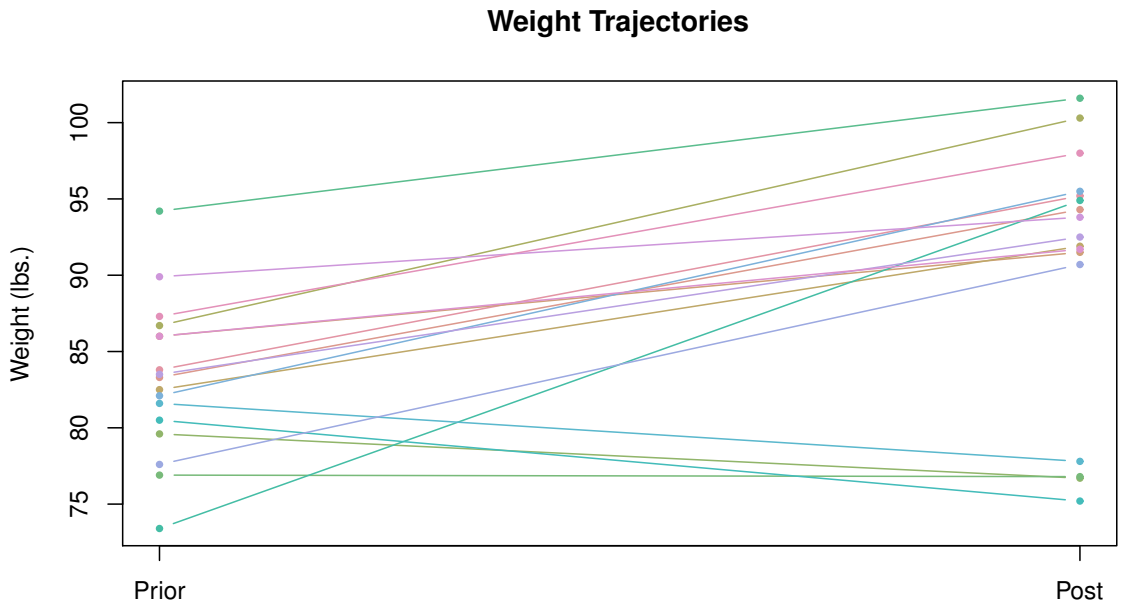

Figure 6: Individual weight trajectories of anorexic girls before and after treatment.

wine. Thus, each of the three wines gets one score from each rater. In total, we 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
```

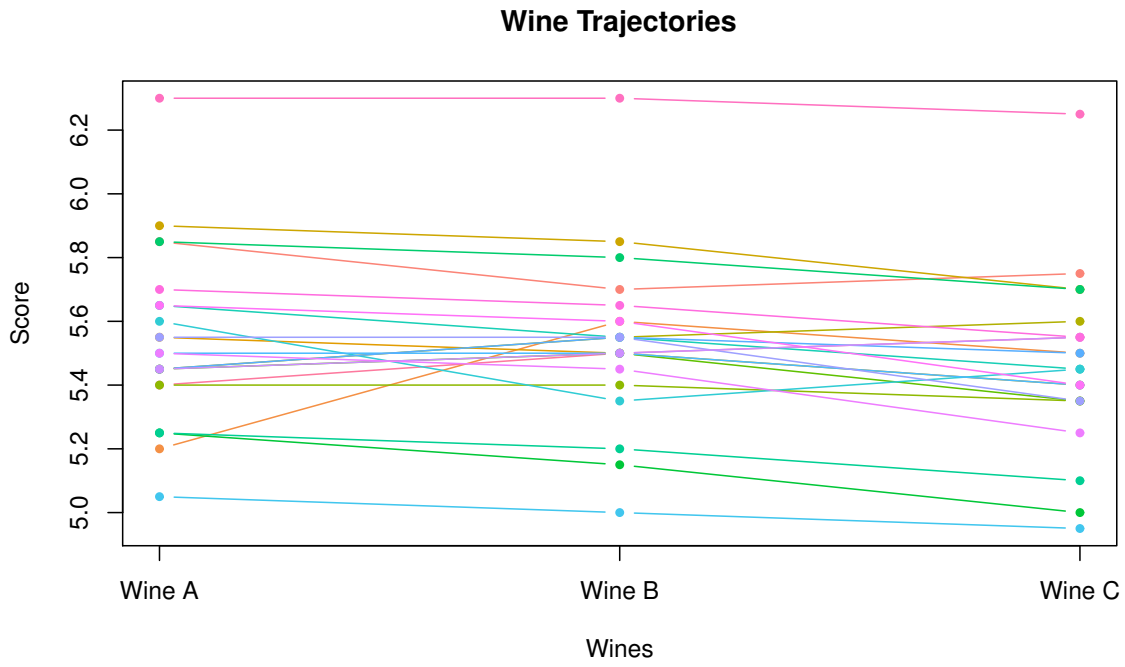

Figure 7: 22 taster trajectories for three different wines.

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 parameteric 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 2013) 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.

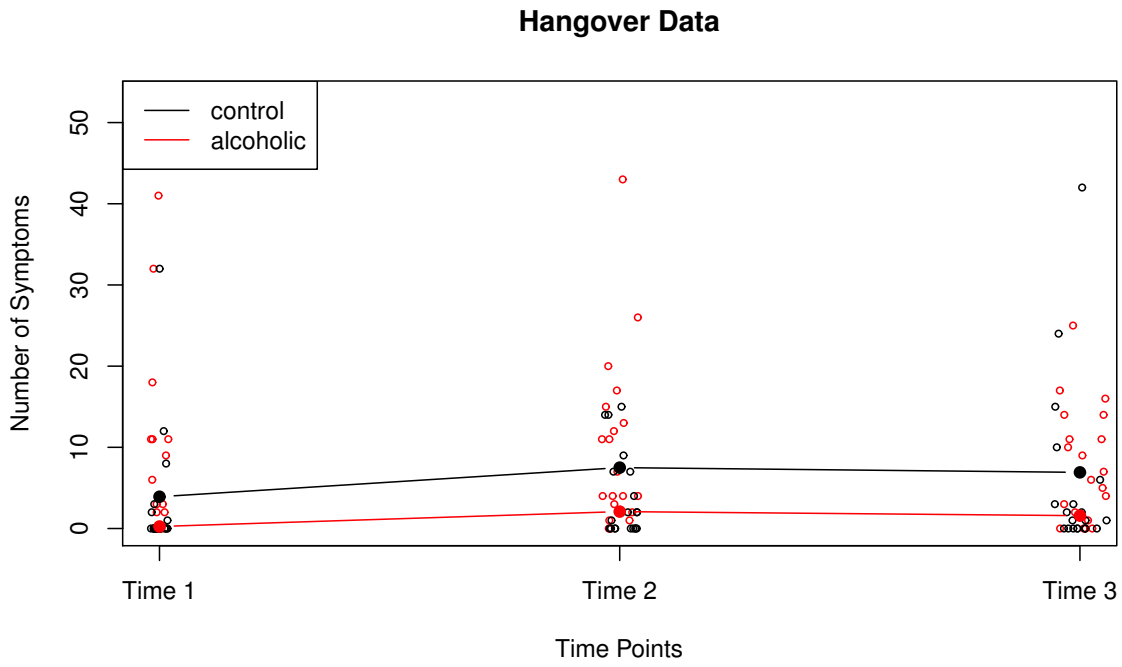

Figure 8: 20% trimmed means of the number of hangover symptoms across three time points.

The 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. For comparison we fit the same model through `ezANOVA` and see that both functions lead to the same results.

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

## Call:
## 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)

## Call:
## sppba(formula = symptoms ~ group * time, id = id, data = hangover)
##
## Test statistic: 4.714
## p-value: 0.024

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

## Call:
## sppbb(formula = symptoms ~ group * time, id = id, data = hangover)
##
## Test statistic: -1.8387 -0.6875 -0.1176
## p-value: 0.118

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

## Call:
## sppbi(formula = symptoms ~ group * time, id = id, data = hangover)
##
## Test statistic: -0.9375 0.4157 -0.5
## p-value: 0.838
```

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

4. Robust Nonparametric ANCOVA

4.1. Running Interval Smoothers

Before we talk about robust ANCOVA, we need to do some elaborations on smoothers. In general, a smoother is a function that approximates the data points while leaving out noise in the data. Smoothing functions typically have a smoothing parameter by 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. The smoothers presneted 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 γ-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., MAD = median$|x_i - \widetilde{x}|$. Let MADN = 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

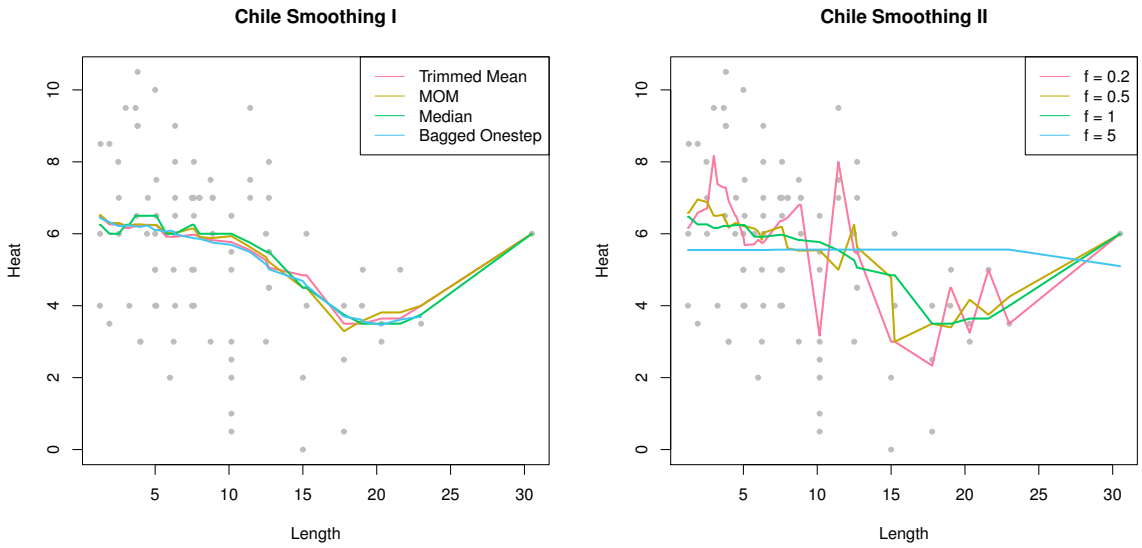

Figure 9: Left panel: smoothers with various robust location measures. Right panel: trimmed mean smoother with varying smoothing parameter f.

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` performs 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 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. It the design point argument (`pts`) is not specified, the routine picks five points automatically (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 are covariate and are interested in possible differences between treatment and control group with respect to the postest 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)
##
##               n: control n: treatment trimmed mean diff     se lower CI
## Pretest = 18          21           20          -11.1128 4.2694 -23.3621
## Pretest = 70          20           21           -3.2186 1.9607  -8.8236
## Pretest = 80          24           23           -2.8146 1.7505  -7.7819
## Pretest = 90          24           22           -5.0670 1.3127  -8.7722
## Pretest = 100         28           30           -1.8444 0.9937  -4.6214
## Pretest = 110         24           22           -1.2491 0.8167  -3.5572
##               upper CI statistic p-value
## Pretest = 18    1.1364    2.6029  0.0163
## Pretest = 70    2.3864    1.6416  0.1143
## Pretest = 80    2.1528    1.6079  0.1203
## Pretest = 90   -1.3617    3.8599  0.0006
## Pretest = 100   0.9325    1.8561  0.0729
## Pretest = 110   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. Discussion

Future updates will include the following robust methods: mediator and moderator models, MANOVA, and intraclass correlation. In addition, functions for computing effect sizes will be available.

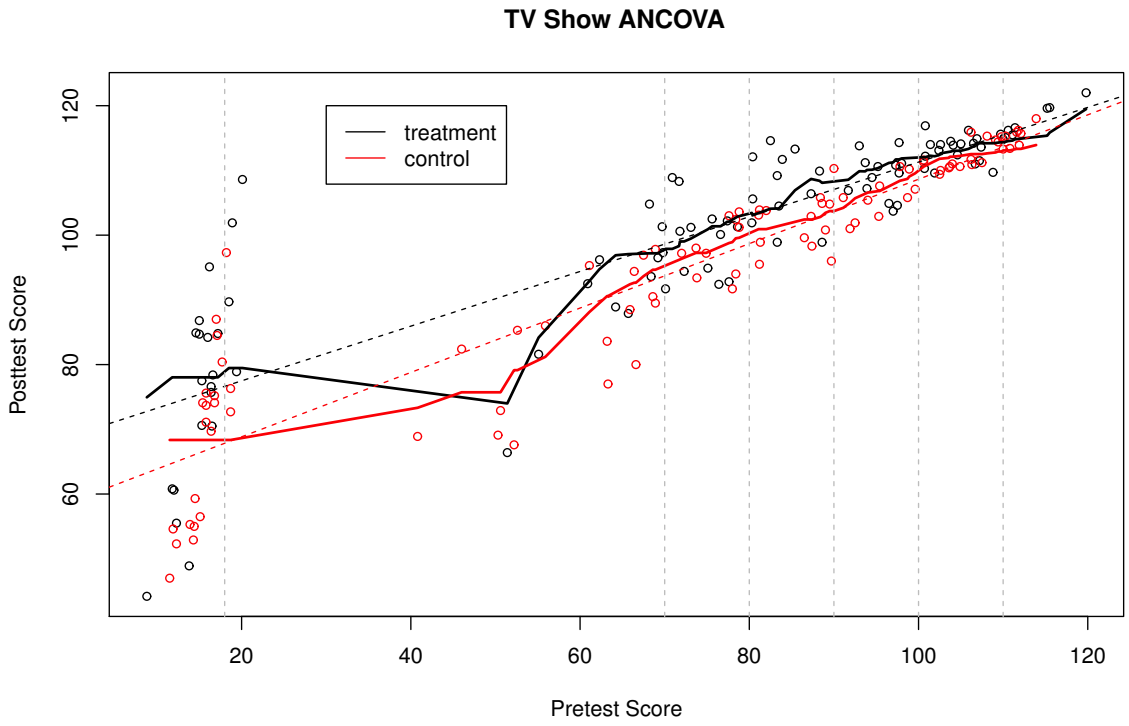

Figure 10: Robust ANCOVA fit on TV show data across treatment and control group. The nonparametric regression lines for both subgroups are shown as well as the OLS fit (dashed lines). The vertical lines show the design points our comparisons are based on.

References

Bates D, Maechler M, Bolker BM, Walker S (2015). "Fitting linear mixed-effects models using **lme4**." *Journal of Statistical Software*. Forthcoming.

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, Series B*, **26**, 211–252.

Canty A, Ripley B (2015). ***boot**: Bootstrap R (S-Plus) functions*. R package version 1.3-17, URL https://cran.r-project.org/web/packages/boot/.

Cressie NAC, Whitford HJ (1986). "How to use the two sample t-test." *Biometrical Journal*, **28**, 131–148.

Field A, Miles J, Field Z (2012). *Discovering Statistics Using R*. Sage, 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.

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. URL http://www.jstatsoft.org/v28/i08/.

Huber PJ (1964). "Robust estimation of location parameters." *Annals of Mathematical Statistics*, **35**, 73–101.

Huber PJ (1981). *Robust Statistics*. Wiley, New York.

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.

Lawrence MA (2013). *ez: Easy analysis and visualization of factorial experiments*. R package version 4.2-2, URL http://CRAN.R-project.org/package=ez.

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.

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.

R Core Team (2015). *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, New York, NY.

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.

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. URL http://www.jstatsoft.org/v32/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. URL http://www.jstatsoft.org/v53/i03/.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer, 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.

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, London, UK.

Yuen KK (1974). "The two sample trimmed t for unequal population variances." *Biometrika*, **61**, 165–170.

Affiliation:

Patrick Mair
Department of Psychology
Harvard University
E-mail: mair@fas.harvard.edu
URL: http://http://scholar.harvard.edu/mair