Robust Statistical Methods Using WRS2

Patrick Mair
Harvard University

Rand Wilcox
University of Southern California

Abstract

This vignette is a (slightly) modified version of Mair and Wilcox (2020), published in Behavior Research Methods.

It introduces the R package WRS2 that implements various robust statistical methods. It elaborates on the basics of robust statistics by introducing robust location, dispersion, and correlation measures. The location and dispersion measures are then used in robust variants of independent and dependent samples t-tests and ANOVA, including between-within subject designs and quantile ANOVA. Further, robust ANCOVA as well as robust mediation models are introduced. The paper targets applied researchers; it is therefore kept rather non-technical and written in a tutorial style. Special emphasis is placed on applications in the social and behavioral sciences and illustrations of how to perform corresponding robust analyses in R. The R code for reproducing the results in the paper is given in the supplementary materials.

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

1. Introduction

Classic inferential methods based on means (e.g., the ANOVA F-test) assume normality and homoscedasticity (equal variances). A fundamental issue is whether violating these two assumptions is a serious practical concern. Based on numerous articles summarized in Wilcox (2017), the answer is an unequivocal "yes". Under general conditions they can have relatively poor power, they can yield inaccurate confidence intervals, and they can poorly characterize the extent groups differ. Even a small departure from normality can be a serious concern. Despite the central limit theorem, certain serious concerns persist even when dealing with large sample sizes. Least squares regression inherits all of these concerns and new concerns are introduced.

A strategy for salvaging classic methods is to test assumptions. For example, test the hypothesis that groups have equal variances and if it fails to reject, assume homoscedasticity. However, published papers summarized in Wilcox (2017) indicate that this strategy is unsatisfactory. Roughly, such tests do not have enough power to detect situations where violating assumptions is a serious practical concern. A simple transformation of the data is another strategy that is unsatisfactory under general conditions.

The family of robust statistical methods offers an attractive framework for dealing with these issues. In some situations robust methods make little practical difference, but they can substantially alter our understanding of data. The only known method for determining whether

this is the case is to simply use a robust method and compare to the results based on a classic technique.

The R (R Core Team 2019) package ecosystem gives the user many possibilities to apply robust methods. A general overview of corresponding implementations is given on the CRAN task view on robust statistics[1]. Here we focus on the **WRS2** package, available on CRAN, that implements methods from the original **WRS** package (Wilcox and Schönbrodt 2017). **WRS2** is less comprehensive than **WRS** but implements the most important functionalities in a user-friendly manner (it uses data frames as basic input structures instead of lists, formula objects for model specification, basic S3 print/summary/plot methods, etc). Here we elaborate on basic data analysis strategies implemented in **WRS2** and especially relevant for the social and behavioral sciences. The article starts with simple robust measures of location, dispersion and correlation, followed by robust group comparison strategies such as t-tests, ANOVA, between-within subject designs, and quantile comparisons. Subsequently, we present robust ANCOVA and robust mediation strategies.

Note that in the text we only give a minimum of technical details, necessary to have a basic understanding of the respective method. An excellent introduction to robust methods within a psychology context is given in Field and Wilcox (2017), more comprehensive treatments are given in Wilcox (2017).

2. Robust Measures of Location, Scale, and Dependence

2.1. Robust Location Measures

A robust alternative to the arithmetic mean $\bar{x}$ is the class of *trimmed means*, which contains the sample median as a special case. A trimmed mean discards a certain percentage at both ends of the distribution. For instance, a 10% trimmed mean cuts off 10% at the lower end and 10% the higher end of the distribution. Let $x_1, \dots x_{10}$ be $n = 10$ sample values, sorted in ascending order. The 10% trimmed sample mean is

$$\bar{x}_t = (x_2 + x_3 + \cdots + x_8 + x_9)/8. \tag{1}$$

That is, it excludes the lowest and the largest value and computes the arithmetic mean on the remaining values. The sample size h after trimming is called effective sample size (here, $h = 8$). Note that if the trimming portion is set to $\gamma = 0.5$, the trimmed mean $\bar{x}_t$ results in the median $\tilde{x}$. An appeal of a 20% trimmed mean is that it achieves nearly the same amount of power as the mean when sampling from a normal distribution. And when there are outliers, a 20% trimmed mean can have a subtantially smaller standard error.

In R, a trimmed mean can be computed via the basic `mean` function by setting the `trim` argument accordingly. Let us illustrate its computation using a simple data vector taken from a self-awareness and self-evaluation study by Dana (1990). The variable reflects the time (in sec.) persons could keep a portion of an apparatus in contact with a specified target. Note that this variable is skewed, which is the standard for duration data. The 10% trimmed mean including the standard error (see Appendix for details) can be computed as follows. For comparison we also report the standard arithmetic mean and its standard error.

[1] URL: http://cran.r-project.org/web/views/Robust.html

```
R> library("WRS2")
R> timevec <- c(77, 87, 88, 114, 151, 210, 219, 246, 253, 262, 296, 299,
+               306, 376, 428, 515, 666, 1310, 2611)
R> mean(timevec, 0.1)

[1] 342.7059

R> trimse(timevec, 0.1)

[1] 103.2686

R> mean(timevec)

[1] 448.1053

R> sd(timevec)/sqrt(length(timevec))

[1] 136.4174
```

The median including standard error from WRS2 is:

```
R> median(timevec)

[1] 262

R> msmedse(timevec)

[1] 77.83901
```

Note that in the case of ties, extant methods for estimating the standard error of the sample median can be highly inaccurate. This includes the method used by `msmedse`. Inferential methods based on a percentile bootstrap effectively deal with this issue, as implemented in the `onesampb` function.

Another robust location alternative to the mean is the *Winsorized mean*. A 10% Winsorized mean, for example, is computed as follows. Rather than discarding the lowest 10% of the values, as done by the 10% trimmed mean, they are set equal to the smallest value not trimmed. In a similar manner, the largest 10% are set equal to the largest value not trimmed. This process is called *Winsorizing*, which in effect transforms the tails of the distribution. Instead of Eq. (1), the 10% Winsorized sample mean uses

$$\bar{x}_w = (x_2 + x_2 + x_3 + \cdots + x_8 + x_9 + x_9)/10. \tag{2}$$

Thus, it replaces the lowest and the largest values by its neighbors and computes the arithmetic mean on this new sequence of values. Similar to the trimmed mean, the amount of Winsorizing (i.e., the *Winsorizing level* γ) has to be chosen *a priori*. The **WRS2** function to compute Winsorized mean is called `winmean`, whereas `winvar` calculates the Winsorized variance.

```
R> winmean(timevec, 0.1)

[1] 380.1579

R> winse(timevec, 0.1)

[1] 92.9417

R> winvar(timevec, 0.1)

[1] 129679
```

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 consider 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 quadratic loss we can think of a more general, differentiable distance function $\xi(\cdot)$:

$$\sum_{i=1}^{n} \xi(x_i - \mu_m) \rightarrow \min! \tag{3}$$

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. Huber (1981), for instance, proposed the following function:

$$\Psi(x) = \begin{cases} x & \text{if } |x| \leq K \\ K\text{sign}(x) & \text{if } |x| > K \end{cases} \tag{4}$$

K is the *bending constant* for which Huber suggested a value of $K = 1.28$. Increasing K decreases sensitivity to the tails of the distribution. The estimation of M-estimators is performed iteratively (see Wilcox 2017, for details) and implemented in the `mest` function.

```
R> mest(timevec)

[1] 285.1576

R> mestse(timevec)

[1] 52.59286
```

Other M-estimators are the one-step estimator and the modified one-step estimator (MOM), as implemented in the functions `onestep` and `mom`. In effect, they empirically determine which values are outliers and eliminate them. One-sample tests for the median, one-step, and MOM are implemented in `onesampb` (using a percentile bootstrap approach). Further details on these measures including expressions for standard errors can be found in Wilcox (2017, Chapter 3).

2.2. Robust Correlation Coefficients

Pearson's correlation is not robust. Outliers can mask a strong association among the bulk of the data and even a slight departure from normality can render it meaningless (Wilcox 2017). Here we present two M-measures of correlation, meaning that they guard against the deleterious impact of outliers among the marginal distributions. The first is the *percentage bend correlation* ρ_{pb}, a robust measure of the linear association between two random variables. When the underlying data are bivariate normal, ρ_{pb} gives essentially the same values as Pearson's ρ. In general, ρ_{pb} is more robust to slight changes in the data than ρ. The computation, involving a bending constant β ($0 \leq \beta \leq 0.5$), is given in Wilcox (2017, p. 491). WRS2 provides the `pbcor` function to calculate the percentage bend correlation coefficient and to perform a one-sample test (H_0: $\rho_{pb} = 0$). For simultaneous inference on a correlation matrix, `pball` can be used. It also includes a statistic H which tests the global hypothesis that all percentage bend correlations in the matrix are equal to 0 in the population.

A second robust correlation measure is the *Winsorized correlation* ρ_w, which requires the specification of the amount of Winsorization. The computation is simple: it uses Person's correlation formula applied on the Winsorized data. 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 Wilcox (2017, p. 452). 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 one was a control group. Here we are interested in the Winsorized correlations across the three time points for the participants in the alcoholic group. The corresponding data subset needs to be organized in wide format with the test occasions in separate columns.

```
R> library("reshape")
R> hangctr <- subset(hangover, subset = group == "alcoholic")
R> hangwide <- cast(hangctr, id ~ time, value = "symptoms")[,-1]
R> colnames(hangwide) <- paste("Time", 1:3)
R> winall(hangwide)


Call:
winall(x = hangwide)

Robust correlation matrix:
       Time 1 Time 2 Time 3
Time 1 1.0000 0.2651 0.4875
Time 2 0.2651 1.0000 0.6791
Time 3 0.4875 0.6791 1.0000

p-values:
        Time 1  Time 2  Time 3
Time 1      NA 0.27046 0.03935
Time 2 0.27046      NA 0.00284
Time 3 0.03935 0.00284      NA
```

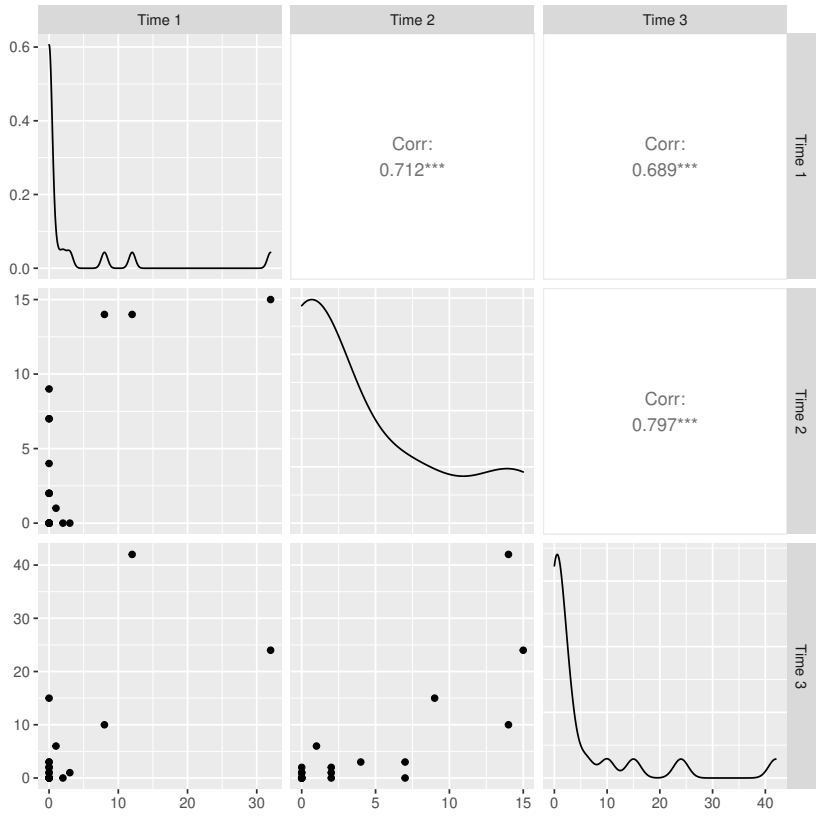

Figure 1: Scatterplot matrix for hangover data. The upper triangle panels report the Pearson correlations.

Figure 1 shows the scatterplot matrix with the Pearson correlations in the upper triangle panels. These correlations clearly differ from the robust correlations reported above.

In order to test for equality of two correlation coefficients, `twopcor` can be used for Pearson correlations and `twocor` for percentage bend or Winsorized correlations. As an example, using the hangover dataset we want to test whether the time 1/time 2 correlation ρ_{pb1} of the control group is the same as the time1/time2 correlation ρ_{pb2} of the alcoholic group.

```
R> ct1 <- subset(hangover, subset = (group == "control" & time == 1))$symp
R> ct2 <- subset(hangover, subset = (group == "control" & time == 2))$symp
R> at1 <- subset(hangover, subset = (group == "alcoholic" & time == 1))$symp
R> at2 <- subset(hangover, subset = (group == "alcoholic" & time == 2))$symp
R> set.seed(123)
R> 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: 0.5886
```

```
Second correlation coefficient: 0.5628
Confidence interval (difference): -0.5783 0.8399
p-value: 0.9219
```

Note that the confidence interval (CI) for the correlation differences is bootstrapped. Other types of robust correlation measures are the well-known Kendall's τ and Spearman's ρ as implemented in the base R `cor` function.

3. Robust Two-Sample Testing Strategies

3.1. Robust Tests for Two Independent Groups and Effect Sizes

Yuen (1974) proposed a test statistic for a two-sample trimmed mean test which allows for the presence of unequal variances. The test statistic is

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2}}, \tag{5}$$

where d_j is an estimate of the squared standard error for $\bar{X}_{tj}$, which is based in part on the Winsorized data. Under the null (H_0: $\mu_{t1} = \mu_{t2}$), the test statistic follows, approximately, a t-distribution[2] with ν_y degrees of freedom (df). Formal expressions for the standard error in Eq. (5) and the df can be found in the Appendix. Note that if no trimming is involved, this method reduces to Welch's classical t-test with unequal variances (Welch 1938), as implemented in `t.test`.

Yuen's test is implemented in the `yuen` function. There is also a bootstrap version (see `yuenbt`) which is suggested to be used when the amount of trimming is close to zero. The example dataset, included in the **WRS2** package, consists of various soccer team statistics in five different European leagues, collected at the end of the 2008/2009 season. Here we focus on the Spanish Primera División (20 teams) and the German Bundesliga (18 teams). The data are organized in an object called `SpainGer` in which the goals scored per game are in the (metric) variable called `GoalsGame`, and the variable `League` is a factor (nominal) indicating whether the team was from the Spanish Primera División or the German Bundesliga.

We are interested in comparing the trimmed means of goals scored per game across these two leagues. The group-wise boxplots with superimposed 1D scatterplots (points jittered) in Figure 2 visualize potential differences in the distributions. Spain has a considerably right-skewed goal distribution involving three outliers (Barcelona, Real Madrid, Atletico Madrid). In the German league, the distribution looks fairly symmetric.

Yuen's test based on the trimmed means with default trimming level of $\gamma = 0.2$ can be computed as follows

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

Call:
yuen(formula = GoalsGame ~ League, data = SpainGer)
```

[2]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.

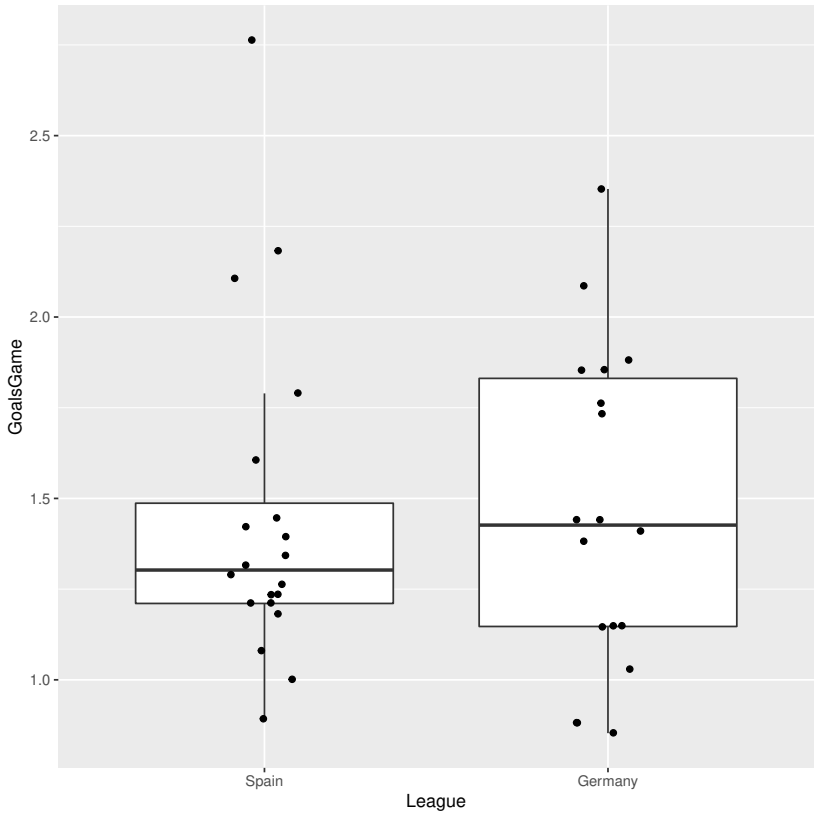

Figure 2: Boxplots for scored goals per game (Spanish vs. German league) with superimposed 1D jittered scatterplot.

```
Test statistic: 0.8394 (df = 16.17), p-value = 0.4135

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

Explanatory measure of effect size: 0.15
```

The 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) propose a robust version of Cohen's d (Cohen 1988).

$$\delta_t = 0.642 \frac{\bar{X}_{t1} - \bar{X}_{t2}}{S_w^*} \tag{6}$$

The formal expression for S_w^* as well as a modification for unequal variances can be found in the Appendix. In WRS2 this effect size (equal variances assumed) can be computed as follows:

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

$AKPeffect
[1] -0.281395

$AKPci
[1] -1.2875813  0.3477103

$alpha
[1] 0.05

$call
akp.effect(formula = GoalsGame ~ League, data = SpainGer)

attr(,"class")
[1] "AKP"
```

The same rules of thumb as for Cohen's d can be used; that is, $|\delta_t| = 0.2$, 0.5, and 0.8 correspond to small, medium, and large effects. However, we would like to point out that these rules should not be used blindly. As Cohen (1988, p. 79) puts it, if one finds that, "what is here defined as large is too small (or too large) to meet what his area of behavioral science would consider appropriate standards is urged to make more suitable operational definitions".

Wilcox and Tian (2011) proposed an *explanatory measure of effect size* ξ which does not require equal variances and can be generalized to multiple group settings. A simple way to introduce this measure is to use the concept of explanatory power from regression with response Y and fitted values $\hat{Y}$:

$$\xi^2 = \frac{\sigma^2(\hat{Y})}{\sigma^2(Y)}, \tag{7}$$

where $\sigma^2(Y)$ is some measure of variation associated with Y. When $\sigma^2(Y)$ is taken to be the usual variance, $\xi^2 = \rho^2$, where ρ is Pearson's correlation.

In a t-test setting with equal samples sizes, $\sigma^2(Y)$ can be simply estimated by the sample variance based on the $2n$ pooled observations, whereas $\sigma^2(\hat{Y})$ is estimated with

$$(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2, \tag{8}$$

where $\bar{X}$ is the grand mean[3]. The explanatory measure of effect size is simply $\xi = \sqrt{\xi^2}$. To make this effect size measure "robust", all that needs to be done is to replace the grand mean $\bar{X}$ and group means $\bar{X}_1$ and $\bar{X}_2$ in Eq. (8) with a robust location measure (e.g., trimmed mean, Winsorized mean, median) in order to estimate $\sigma^2(\hat{Y})$. The variance $\sigma^2(Y)$ needs to be replaced by the corresponding robust variance estimator (e.g., Winsorized variance).

In **WRS2**, the explanatory measure of effect size can be computed as follows:

```
R> set.seed(123)
R> yuen.effect.ci(GoalsGame ~ League, data = SpainGer)

$effsize
[1] 0.1506268

$alpha
[1] 0.05

$CI
[1] 0.0000000 0.6134423
```

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 (CI) for $\hat{\xi}$ based on a percentile bootstrap. Varying dispersions in the response variable across the factor levels (*heteroscedasticity*) are allowed.

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

```
R> set.seed(123)
R> pb2gen(GoalsGame ~ League, data = SpainGer, est = "median")

Call:
pb2gen(formula = GoalsGame ~ League, data = SpainGer, est = "median")

Test statistic: -0.1238, p-value = 0.39733
95% confidence interval:
-0.5062    0.2214

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

[3]For unequal sample sizes a modified estimator is used that accounts for unbalancedness in the data.

```
Call:
pb2gen(formula = GoalsGame ~ League, data = SpainGer, est = "onestep")

Test statistic: -0.1181, p-value = 0.39065
95% confidence interval:
-0.3838    0.1894
```

These tests simply use the differences in medians (i.e., $\tilde{X}_1 - \tilde{X}_2$) and differences in Huber's Ψ estimator from Eq. (4) (i.e., $\Psi(X_1) - \Psi(X_2)$), respectively, as test statistics. CIs and p-values are determined through bootstrap. Currently, when using the median and there are tied values, this is the only known method that performs well in simulations (Wilcox 2017).

Another function implemented in **WRS2** is `qcomhd` for general quantile comparison across two groups (Wilcox, Erceg-Hurn, Clark, and Carlson 2014) using the quantile estimator proposed by Harrell and Davis (1982). The null hypothesis is simply H_0: $\theta_{q1} = \theta_{q2}$, where θ_{q1} and θ_{q2} are the q-th quantiles in group 1 and 2, respectively. Confidence intervals for $\hat{\theta}_{q1} - \hat{\theta}_{q2}$ and p-values are determined via a percentile bootstrap. This test provides a more detailed understanding of where and how distributions differ. Let us apply this approach on the same data as above. We keep the default setting which tests for differences in the 0.1, 0.25, 0.5, 0.75, and 0.95 quantiles. Note that the sample size is slightly small to apply this test[4].

```
R> set.seed(123)
R> fitqt <- qcomhd(GoalsGame ~ League, data = SpainGer,
+   q = c(0.1, 0.25, 0.5, 0.75, 0.95), nboot = 500)
R> 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 -0.1765 0.3259 0.0100   0.232
2 0.25 20 18 1.1950 1.0892     0.1058 -0.2335 0.3132 0.0167   0.436
3 0.50 20 18 1.3109 1.4304    -0.1194 -0.4656 0.2571 0.0125   0.492
4 0.75 20 18 1.6220 1.8078    -0.1858 -0.5377 0.3983 0.0500   0.524
5 0.95 20 18 2.5160 2.2402     0.2758 -0.6529 0.8150 0.0250   0.556
```

The p-values are adjusted using Hochberg's method[5] (see `p.crit` for the critical values the p-values in the last column should be compared to). Note that ties in the data are not problematic for this particular test. Plots that illustrate the results of quantile difference tests are implemented in the **rogme** package (Rousselet, Pernet, and Wilcox 2017).

3.2. Robust Tests for Two Dependent Groups

Yuen's trimmed mean t-test in Eq. (5) can be generalized to paired sample settings as follows:

[4] It is suggested to have at least 20 observations in each group.
[5] A brief explanation of Hochberg's method can be found in the Appendix.

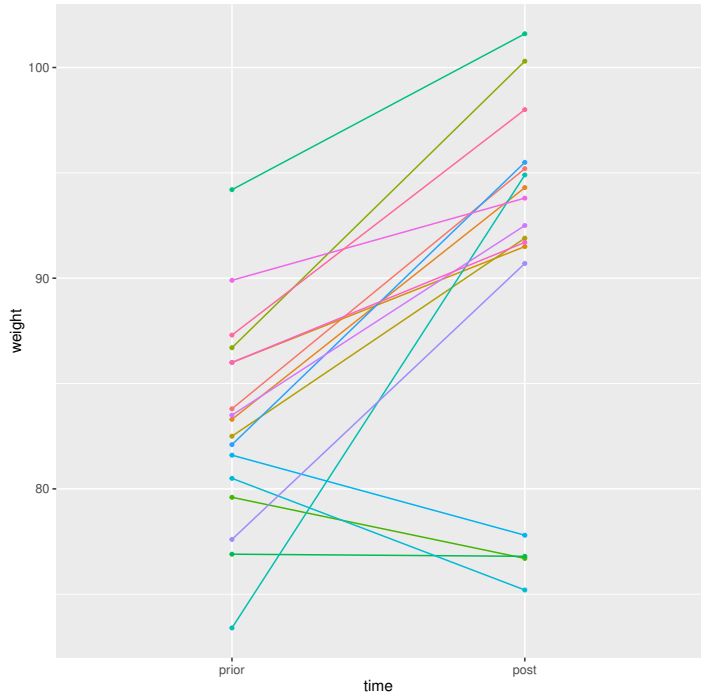

Figure 3: Individual weight trajectories of anorexic girls before and after treatment.

$$T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2 + d_{12}}} \tag{9}$$

Under the null (H_0: $\mu_{t1} = \mu_{t2}$), T_y is t-distributed with $df = h - 1$, where h is the effective sample size. Details on the computation of this statistic can be found in the Appendix.

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 (Venables and Ripley 2002) and presents data pairs involving weights of girls before and after treatment for anorexia. We use a subset of 17 girls from the family treatment (`FT`) condition. Figure 3 presents the individual trajectories. We keep the default trimming level (20%) and get the following test results.

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

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
```

The output suggests that overall the treatment was successful. The explanatory measure of effect size, constructed according to the same principles as outlined above, suggests a large effect.

Quantile comparisons for paired samples (H_0: $\theta_{q1} = \theta_{q2}$) can be computed using `Dqcomhd` (Wilcox and Erceg-Hurn 2012). As the independent sample version in `qcomhd`, it uses the quantile estimator proposed by Harrell and Davis (1982), and bootstrapping to determine the CI for $\hat{\theta}_{q1} - \hat{\theta}_{q2}$ and the p-values (corrected for multiple testing).

```
R> set.seed(123)
R> with(anorexiaFT, Dqcomhd(Prewt, Postwt, q = c(0.25, 0.5, 0.75)))

Call:
Dqcomhd(x = Prewt, y = Postwt, q = c(0.25, 0.5, 0.75))

Parameter table:
     q n1 n2    est1    est2 est1-est.2   ci.low   ci.up p.crit p.value
1 0.25 17 17 79.9588 84.5667    -4.6079 -12.3789  2.4284 0.0500   0.270
2 0.50 17 17 83.1703 92.7727    -9.6024 -11.8158 -4.7773 0.0250   0.004
3 0.75 17 17 86.3380 95.8962    -9.5583 -11.9286 -6.9418 0.0167   0.000
```

We obtain significant weight decrease effects for the second and the third weight quartiles, but not for the first quartile.

3.3. Comparing Two Discrete Distributions

Having two discrete variables X and Y (small sample space), sometimes it is 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). The test statistic is given in the Appendix.

Let us look at a simple artificial example involving responses on a five-point rating scale item across two groups of participants with group sizes n_1 and n_2. The `binband` function compares the two distributions at each possible value (here $1, 2, \ldots, 5$) in the joint sample space.

```
R> g1 <- c(2, 4, 4, 2, 2, 2, 4, 3, 2, 4, 2, 3, 2, 4, 3, 2, 2, 3, 5, 5, 2, 2)
R> g2 <- c(5, 1, 4, 4, 2, 3, 3, 1, 1, 1, 1, 2, 2, 1, 1, 5, 3, 5)
R> binband(g1, g2, KMS = TRUE)

Call:
binband(x = g1, y = g2, 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
```

The CIs are determined using the Kulinskaya-Morgenthaler-Staudte method (`KMS = TRUE`). The function uses Hochberg's multiple comparison adjustment to determine critical p-values with the goal of controlling the probability of one or more Type I errors. The results suggest that the distributions differ significantly at $(x, y) = 1$ only ($p \leq p_{crit}$).

4. One-Way Robust Testing Strategies

Often it is said that F-tests are quite robust against normality violations. As Field and Wilcox (2017, p. 37) recommend, such statements should be banned because based on many papers published during the past fifty years, it is well established that this statement is not correct (especially when dealing with heavy-tailed distributions, unequal sample sizes, and distributions differing in skewness). In this section we present various robust one-way ANOVA strategies, followed by higher order models in the next section.

4.1. One-Way Trimmed Means Comparisons

The first robust ANOVA alternative presented here is a one-way comparison of J trimmed group means ($H_0 : \mu_{t1} = \mu_{t2} = \cdots = \mu_{tJ}$), allowing for heteroscedasticity. Technical details on this F-distributed Welch-type test statistic (Welch 1951) can be found in the Appendix.

In **WRS2** this approach is implemented via the `t1way` function, here applied to the weight differences in the anorexia data from above (post-treatment weight minus pre-treatment weight, resulting in metric variable `Wdiff`). There are two different types of treatment in the data (family treatment `FT` and cognitive behavioral treatment `CBT`) as well as one control group, specified in the factor `Treat`. Figure 4 shows the corresponding boxplots with superimposed 1D scatterplots.

The robust one-way ANOVA based on trimmed means (20% trimming level) can be computed as follows:

```
R> anorexia$Wdiff <- anorexia$Postwt - anorexia$Prewt
R> t1way(Wdiff ~ Treat, data = anorexia)

Call:
t1way(formula = Wdiff ~ Treat, data = anorexia)

Test statistic: F = 5.6286
Degrees of freedom 1: 2
Degrees of freedom 2: 24.89
p-value: 0.00962
```

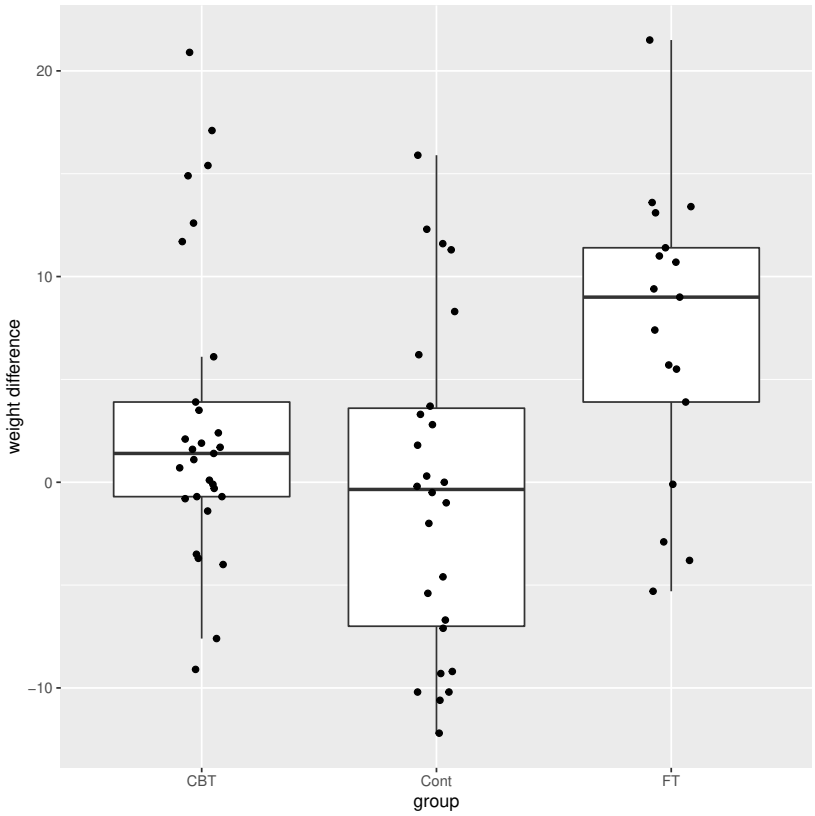

Figure 4: Boxplots with superimposed jittered 1D scatterplots for weight differences across control and two treatment conditions.

```
Explanatory measure of effect size: 0.5
Bootstrap CI: [0.13; 0.77]
```

There is a significant overall effect in weight differences across the treatments. The explanatory measure of effect size ξ follows the same logic as outlined in Eq. (7). The difference compared to the two-sample version is that Eq. (8) generalizes to

$$\sigma^2(\hat{Y}) = \frac{1}{J-1}\sum_{j=1}^{J}(\bar{Y}_j - \bar{Y})^2. \tag{10}$$

The same rules of thumb apply as in the two-sample case. In this example we obtain a large effect.

Post hoc tests on trimmed means use the linear contrast expression

$$\hat{\Psi} = \sum_{j=1}^{J} c_j \bar{X}_{tj}. \tag{11}$$

In WRS2 the constants are specified in a way such that all pairwise post hoc tests are carried out. For instance, for comparing the first two trimmed means $c_1 = 1$ and $c_2 = -1$, whereas the remaining c's are 0.

```
R> lincon(Wdiff ~ Treat, data = anorexia)

Call:
lincon(formula = Wdiff ~ Treat, data = anorexia)

               psihat  ci.lower ci.upper p.value
CBT vs. Cont  2.96250  -3.03709  8.96209 0.22201
CBT vs. FT   -6.10909 -12.33490  0.11672 0.03885
Cont vs. FT  -9.07159 -16.08255 -2.06064 0.00880
```

The function reports the $\hat{\Psi}$ value according to Eq. (11) denoting pairwise trimmed mean differences. The 95% CIs and the p-values are adjusted for multiple testing in the sense that the simultaneous probability coverage of the CIs is $1 - \alpha$ and the family-wise error rate is α. Details on this procedure can be found in Wilcox (1986). A bootstrap version of `t1way` is implemented in `t1waybt` with corresponding bootstrap post hocs in `mcppb20`.

Note that in order to perform linear contrasts, there is no need to first obtain a significant omnibus ANOVA. In many experimental situations, researchers have specific predictions about certain contrasts which can be directly tested (i.e., without computing an omnibus test first).

4.2. One-Way Quantile Comparisons

In this section we focus on testing $H_0 : \theta_1 = \ldots = \theta_J$, where the θ's represent a particular quantile in group j. Let us start with testing for equality of medians across J groups. The test statistic F_M, given in the Appendix, follows the same concept as the one for trimmed means above; the only difference is that it uses an alternative estimate for the standard error. Using our anorexia dataset, it can be computed as follows:

```
R> set.seed(123)
R> med1way(Wdiff ~ Treat, data = anorexia)

Call:
med1way(formula = Wdiff ~ Treat, data = anorexia)

Test statistic F: 4.5708
Critical value: 2.8398
p-value: 0.008
```

A few remarks regarding this test statistic. First, it has been found that by evaluating the test statistic using the df as quoted in the Appendix (i.e., $\nu_1 = J - 1$ and $\nu_2 = \infty$) can result in the actual level being less than the nominal level, (i.e., around 0.02-0.025 when testing at the 0.05 level and n is small). A better strategy, as provided by this implementation, is to simulate the critical value and computing the p-value accordingly. In order to make the result reproducible, above we set a seed.

Second, if there are too many ties in the data, the standard error becomes inaccurate. In such situations, the `Qanova` function provides a good alternative, which allows for general quantile testing across J groups, not only the median. Similar to `qcomhd`, the quantile ANOVA

implemented in `Qanova` uses the Harrel-Davis estimator for the quantiles. It tests the global hypothesis:

$$H_0: \quad \theta_{q1} - \theta_{q2} = \theta_{q2} - \theta_{q3} = \ldots = \theta_{q(J-1)} - \theta_{qJ}.$$

The p-value is determined using a bootstrap (see Wilcox 2017, p. 378–379 for details). In case multiple quantiles are tested at the same time, the p-values are corrected using Hochberg's method.

```
R> set.seed(123)
R> fitqa <- Qanova(Wdiff ~ Treat, data = anorexia,
+    q = c(0.25, 0.5, 0.75))
R> fitqa

Call:
Qanova(formula = Wdiff ~ Treat, data = anorexia, q = c(0.25,
    0.5, 0.75))

         p.value  p.adj
q = 0.25  0.0050 0.0100
q = 0.5   0.0017 0.0050
q = 0.75  0.0417 0.0417
```

It reports the unadjusted and adjusted p-values, to be compared to the α-level. We find significant overall differences at each of the quartiles.

5. Robust Two-Way and Three-Way Comparisons

This section elaborates on higher order ANOVA designs including post hoc tests. Note that all **WRS2** robust ANOVA functions allow the user to fit the full model (i.e., including all possible interactions) only. For more parsimonious models and specific post hoc contrasts, it is suggested to use the corresponding **WRS** functions from Wilcox and Schönbrodt (2017).

5.1. Robust Two-Way ANOVA Strategies

Let us start with a 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, as implemented in `t1way`, can be generalized to two-way designs; details are given in the Appendix. The hypothesis to be tested are the usual two-way ANOVA hypotheses using the trimmed means. Let μ_t be the grand trimmed mean (population), μ_{tjk} the mean in factor level combination jk, $\mu_{tj\cdot}$ the trimmed factor level means of the first factor, and $\mu_{t\cdot k}$ the trimmed factor level means for the second factor. Let $\alpha_j = \mu_{tj\cdot} - \mu_t$, $\beta_k = \mu_{t\cdot k} - \mu_t$, and $(\alpha\beta)_{jk} = \mu_{tjk} - \mu_{tj\cdot} - \mu_{t\cdot k} + \mu_t$. Using this notation, the null hypotheses are:

- First factor: $H_0 : \sum_{j=1}^{J} \alpha_j^2 = 0$.
- Second factor: $H_0 : \sum_{k=1}^{K} \beta_k^2 = 0$.

- Interaction: $H_0 : \sum_{j=1}^{J} \sum_{k=1}^{K} (\alpha\beta)_{jk}^2 = 0$.

Such a robust two-way ANOVA can be carried out using the function `t2way`. To illustrate, we use the beer goggles dataset by Field, Miles, and Field (2012) who studied 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 study we have the 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 5 shows the interaction plots using the trimmed mean (20% trimming level) as location measure. The two-way ANOVA on the trimmed means can be fitted as follows.

```
R> goggles$alcohol <- relevel(goggles$alcohol, ref = "None")
R> 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
```

Not surprisingly, based on what we see in Figure 5, the interaction between gender and alcohol is significant.

Post hoc tests can be applied using the `mcp2atm` function, which, internally calls the `lincon` function described above.

```
R> postgoggle <- mcp2atm(attractiveness ~ gender*alcohol, data = goggles)
R> postgoggle$contrasts

              gender1 alcohol1 alcohol2 alcohol3 gender1:alcohol1
Female_None         1        1        1        0                1
Female_2 Pints      1       -1        0        1               -1
Female_4 Pints      1        0       -1       -1                0
Male_None          -1        1        1        0               -1
Male_2 Pints       -1       -1        0        1                1
Male_4 Pints       -1        0       -1       -1                0
              gender1:alcohol2 gender1:alcohol3
Female_None                  1                0
Female_2 Pints               0                1
Female_4 Pints              -1               -1
Male_None                   -1                0
Male_2 Pints                 0               -1
Male_4 Pints                 1                1
```

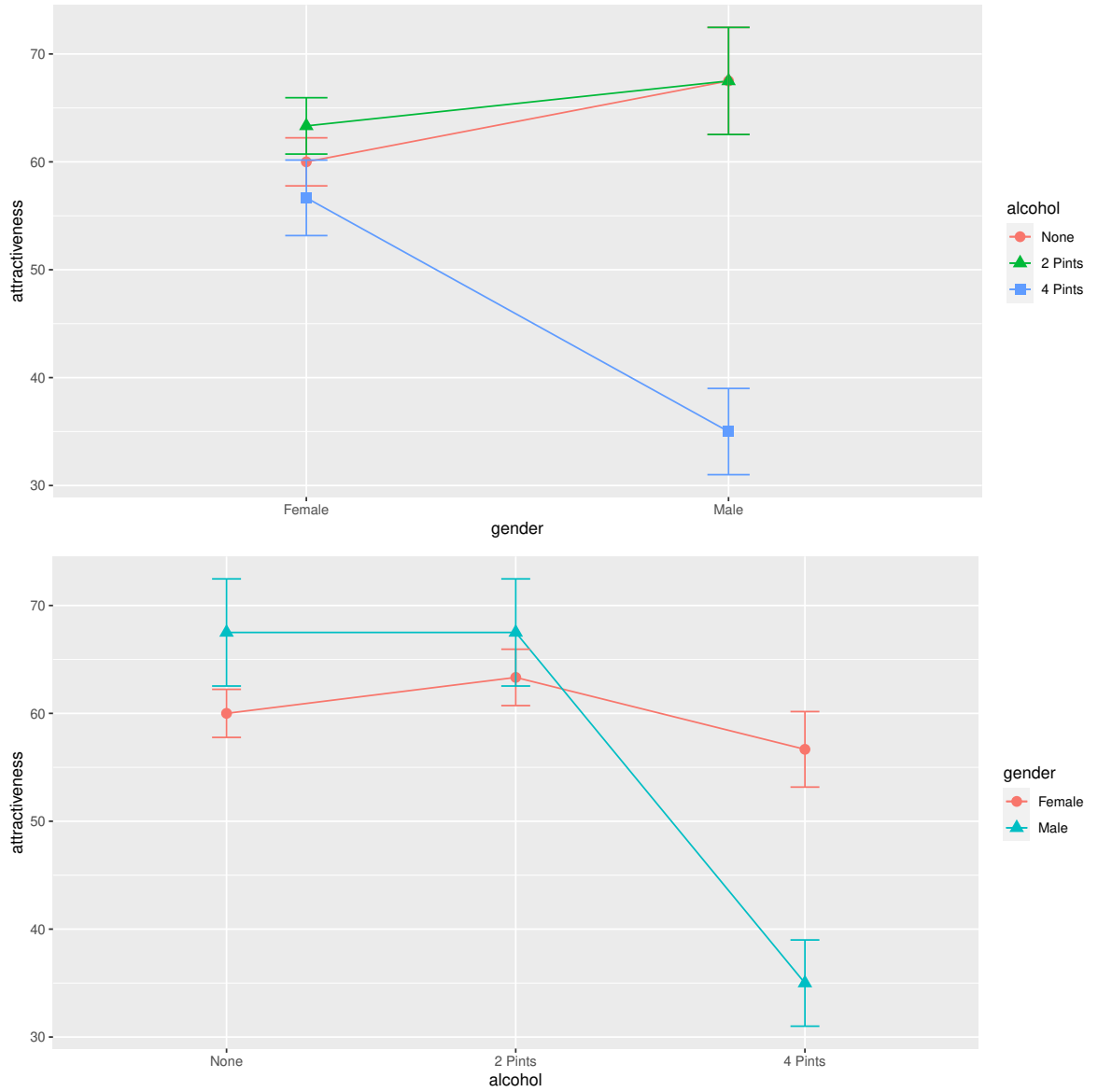

Figure 5: Trimmed means interaction plots for beer goggles dataset.

The second line prints the contrast matrix which illustrates what effects are actually being tested. The results are the following:

```
R> postgoggle

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
gender1:alcohol2 -29.16667 -45.67245 -12.66089 0.00025
gender1:alcohol3 -25.83333 -42.53204  -9.13463 0.00080
```

Let us focus on the interaction first by starting at the bottom. The last effect tells us that the difference attractiveness ratings for 4 pints vs. 2 pints differs significantly in men and women. Similary, the second to last effect tells us that this significant gender difference also applies to 4 pints vs. none. However, males and females do not behave differently if we look at 2 pints vs. none (no significant effect; see third line from the bottom). Note that the 95% CIs and the p-values are adjusted for multiple testing.

Other options for robust two-way ANOVAs are median comparisons using `med2way`, and general M-estimator comparisons using `pbad2way`. For both functions post hoc comparisons can be computed using `mcp2a` (the estimator argument needs to be specified correspondingly) which uses precentile bootstrap for CIs and p-values. Using the beer goggles dataset, the function calls for median and modified one-step estimators (MOM) are the following.

```
R> set.seed(123)
R> med2way(attractiveness ~ gender*alcohol, data = goggles)
R> mcp2a(attractiveness ~ gender*alcohol, data = goggles, est = "median")
R> pbad2way(attractiveness ~ gender*alcohol, data = goggles, est = "mom")
R> mcp2a(attractiveness ~ gender*alcohol, data = goggles, est = "mom")
```

We omit showing the output here; the results are consistent with the trimmed mean comparisons above. Formal details on the median test are given in the Appendix; elaborations on M-estimator comparisons are given in Wilcox (2017, p. 385–388).

5.2. Robust Three-Way ANOVA Strategies

Having three-way designs, **WRS2** provides the function `t3way` for robust ANOVA based on trimmed means. The test statistics are determined according to the same principles as in `t2way` (see Appendix). Again, the critical values are adjusted such that no df of the χ^2-distributed test statistics are reported (see Wilcox 2017, p. 341–346, for details).

The dataset we use to illustrate this approach 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 6 shows two separate interaction plots for male and female swimmers, using the 20% trimmed means.

A three-way robust ANOVA on the trimmed means using `t3way` can be computed as follows:

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

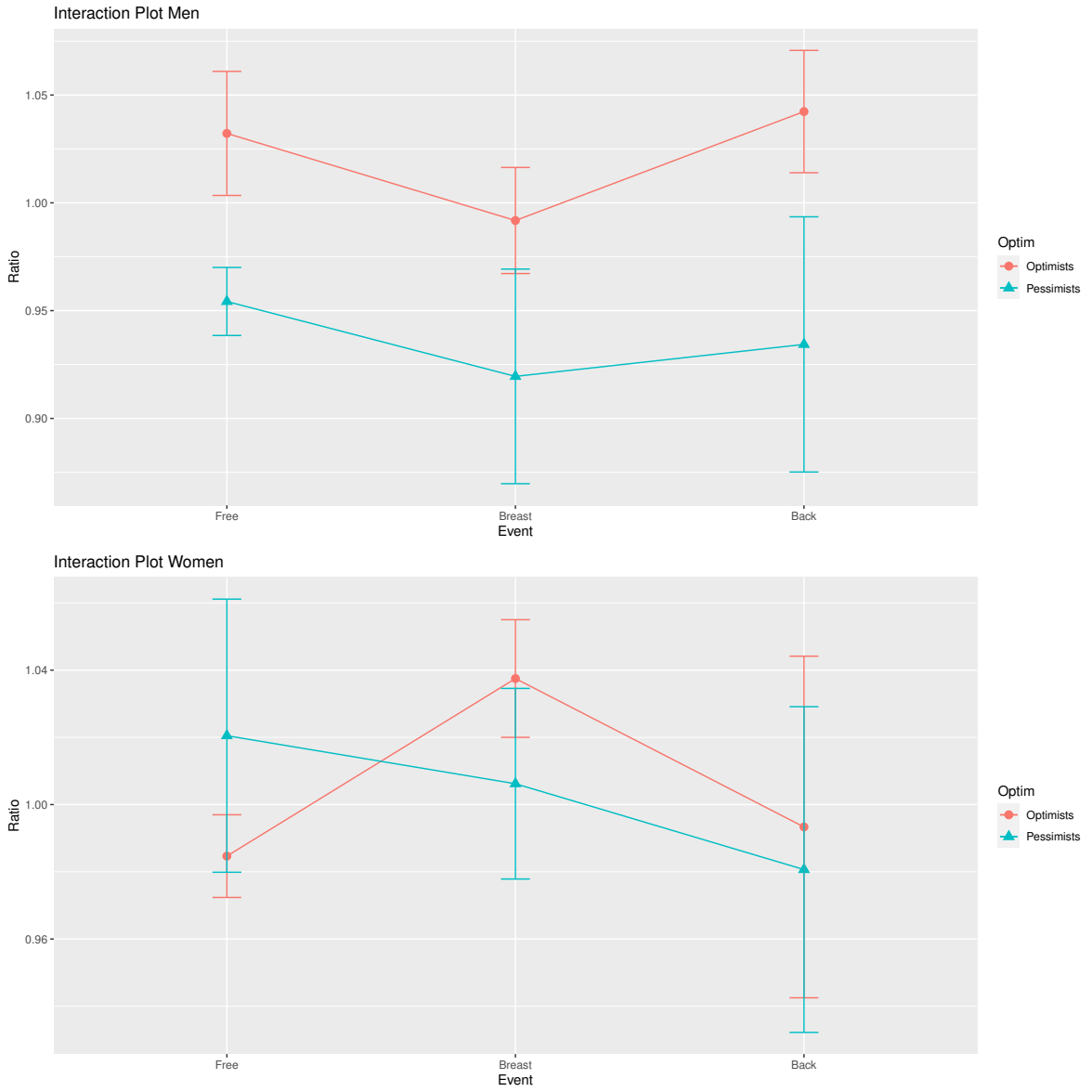

Figure 6: Interaction plot for the trimmed means of the time ratio response for males and females separately.

```
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
```

The crucial effect for interpretation is the significant `Optim:Sex` two-way interaction. We could produce corresponding two-way interaction plots and see that, independently from the swimming style, for the females it does not matter whether someone is an optimist or a pessimist, the time ratio does not change drastically. For the males, there is a substantial difference in the time ratio for optimists and pessimists.

6. Repeated Measurement and Mixed ANOVA Designs

6.1. Paired Samples/Repeated Measurement Designs

In this section we consider paired samples/repeated measurement designs for more than two dependent groups/time points. The WRS2 package provides an implementation of a robust heteroscedastic repeated measurement ANOVA based on the trimmed means. The formulas for the test statistic and the df computations are given in the Appendix.

In **WRS2**, the function to compute a robust repeated measurements ANOVA is `rmanova` with corresponding post hoc tests in `rmmcp`. The data need to be in long format and balanced across the groups. Each of these functions takes three arguments: 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`).

Once more we use the hangover dataset from above, where 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 7.

Here we focus on a single between subjects factor only: control group. In the next section we consider the full dataset with the corresponding between-within subjects design. After subsetting the data accordingly, a robust repeated measurement ANOVA using the `rmanova` function can be fitted as follows:

```
R> hangoverC <- subset(hangover, subset = group == "control")
R> with(hangoverC, rmanova(y = symptoms, groups = time, block = id))

Call:
rmanova(y = symptoms, groups = time, blocks = id)

Test statistic: F = 2.6883
Degrees of freedom 1: 2
Degrees of freedom 2: 22
p-value: 0.09026
```

Post hoc tests (linear contrasts) can be performed as follows:

```
R> with(hangoverC, rmmcp(y = symptoms, groups = time, block = id))
```

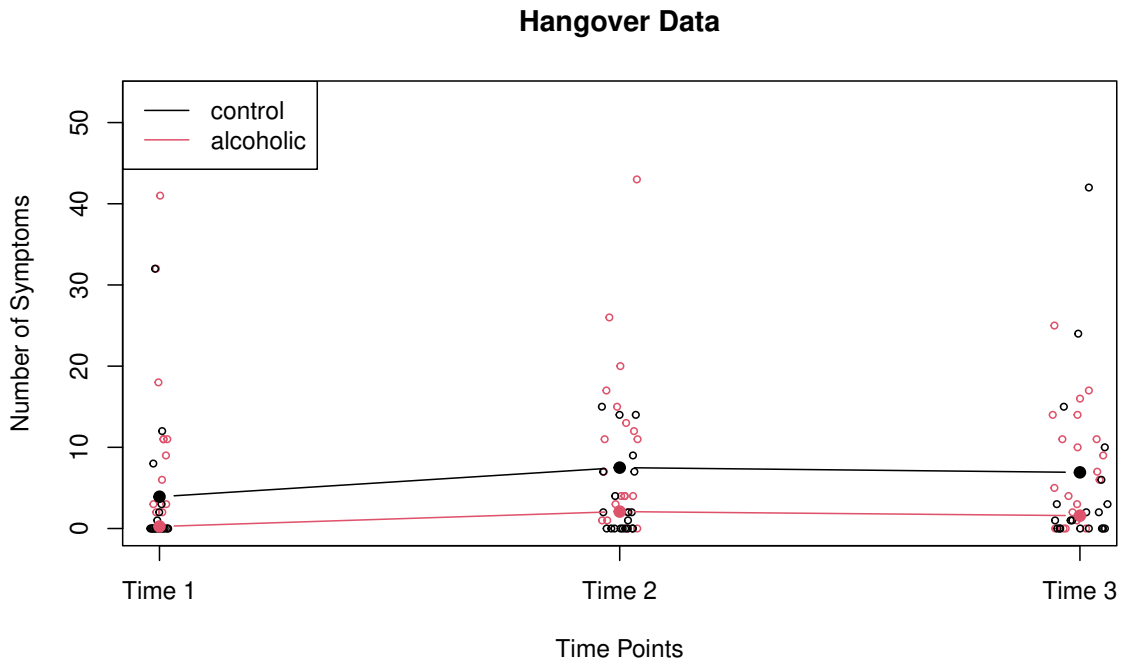

Figure 7: 20% trimmed means of the number of hangover symptoms across three time points.

```
Call:
rmmcp(y = symptoms, groups = time, blocks = id)

          psihat ci.lower ci.upper p.value p.crit   sig
1 vs. 2 -2.66667 -7.47192  2.13858 0.14588 0.0169 FALSE
1 vs. 3 -1.00000 -3.17265  1.17265 0.22085 0.0250 FALSE
2 vs. 3  0.50000 -2.57826  3.57826 0.65583 0.0500 FALSE
```

The `rmmcp` function uses Hochberg's approach to control for the family-wise error (FWE). The bootstrap version of `rmanova` is `rmanovab` with bootstrap post hocs in `pairdepb`.

6.2. Mixed Designs

Let us extend the ANOVA setting above towards mixed designs. That is, we have within-subjects effects (e.g., due to repeated measurements) and between-subjects effects (group comparisons). 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.

We use the hangover data from above and fit a between-within subjects ANOVA on the 20%

trimmed means:

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

Call:
bwtrim(formula = symptoms ~ group * time, id = id, data = hangover)

            value df1     df2 p.value
group      6.6087   1 14.4847  0.0218
time       4.4931   2 15.4173  0.0290
group:time 0.5663   2 15.4173  0.5790
```

We get a non-significant interaction; both main effects are significant.

We can also perform post hoc comparisons on the single effects. WRS2 implements a bootstrap based approach for one-step M estimators, modified one-step estimators (MOM), and medians. To illustrate the hypotheses being tested, we use a different dataset with a slightly more complex design (in terms of the number of factor levels). The study by McGrath (2016) looked at the effects of two forms of written corrective feedback on lexico-grammatical accuracy (`errorRatio`) in the academic writing of English as a foreign language university students. It had a 3×4 within-by-between design with three groups (two treatment and one control; `group`) measured over four occasions (pre-test, treatment, post-test, delayed post-test; `essay`).

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

```
R> set.seed(123)
R> 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 corrresponding population parameters are the same:
p-value: 0.41
```

The p-value suggests that we cannot 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 measurement occasion ($K = 4$). 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 null hypotheses 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:

```
R> set.seed(123)
R> 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 corrresponding population parameters are the same:
p-value: 0.474
```

Again, we cannot reject H_0.

Using this principle, 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

$$\bar{\theta}_{j\cdot} = \frac{1}{K}\sum_{k=1}^{K}\theta_{jk}. \tag{12}$$

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 as follows by using the default `avg = TRUE`:

```
R> set.seed(123)
R> sppba(errorRatio ~ group*essay, id, data = essays)

Call:
sppba(formula = errorRatio ~ group * essay, id = id, data = essays)

Test statistics:
                 Estimate
Control-Indirect   0.2243
Control-Direct     0.1054
Indirect-Direct   -0.1189

Test whether the corrresponding population parameters are the same:
p-value: 0.476
```

Finally, let us elaborate on the `sppbi` function which performs tests on the interactions. In the `sppbb` call six 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 a 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 hypothesis is:

```
R> set.seed(123)
R> sppbi(errorRatio ~ group*essay, id, data = essays)

Call:
sppbi(formula = errorRatio ~ group * essay, id = id, data = essays)

Test statistics:
                               Estimate
essay1-essay2 Control-Indirect -0.14667
essay1-essay2 Control-Direct    0.12083
essay1-essay2 Indirect-Direct   0.26750
essay1-essay3 Control-Indirect -0.11778
essay1-essay3 Control-Direct   -0.02222
essay1-essay3 Indirect-Direct   0.09556
essay1-essay4 Control-Indirect -0.23600
essay1-essay4 Control-Direct    0.21678
essay1-essay4 Indirect-Direct   0.45278
essay2-essay3 Control-Indirect  0.19293
essay2-essay3 Control-Direct   -0.07889
essay2-essay3 Indirect-Direct  -0.27182
essay2-essay4 Control-Indirect  0.10571
essay2-essay4 Control-Direct    0.26905
essay2-essay4 Indirect-Direct   0.16333
essay3-essay4 Control-Indirect -0.20221
essay3-essay4 Control-Direct    0.10643
essay3-essay4 Indirect-Direct   0.30864

Test whether the corrresponding population parameters are the same:
p-value: 0.682
```

Again, we cannot reject H_0.

7. Robust nonparametric ANCOVA

7.1. Running interval smoothers

In this section we introduce a robust ANCOVA version which uses smoothing internally. When dealing with regression, there are situations the usual linear model appears to suffice. But it is well established that parametric regression models can be highly unsatisfactory. 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. The smoothers presented here involve robust location measures from above and are called *running interval smoothers* which work as follows.

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 2017). Let MAD be the *median absolute deviation*, that is, MAD = median$|X_i - \widetilde{X}|$. Let MADN = MAD$/z_{0.75}$, where $z_{0.75}$ represents the 0.75 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 called 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. Let $\hat{\theta}_i$ be a robust location parameter of interest. A running interval smoother computes n $\hat{\theta}_i$ values 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. Then 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 taken from Wright and London (2009) where we have measurements for the length of a chile and its heat (scored on a scale from 0-11). We study various f values and various robust location measures $\hat{\theta}_i$. The left panel in Figure 8 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 various smoothers (see functions `runmean`, `rungen`, and `runmbo`) among the various location measures. However, the choice of the smoothing parameter f affects the function heavily.

7.2. Robust ANCOVA

ANCOVA involves a factorial design and metric covariates that were not part of the experimental manipulation. It assumes homogeneity of regression slopes across the groups when regressing the dependent variable on the covariate. In addition, normality is assumed as well as two types of homoscedasticity. Violating any of these assumptions can have a serious negative impact on the classic ANCOVA method. 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 compared for subgroup differences at various points of interest along the x-continuum.

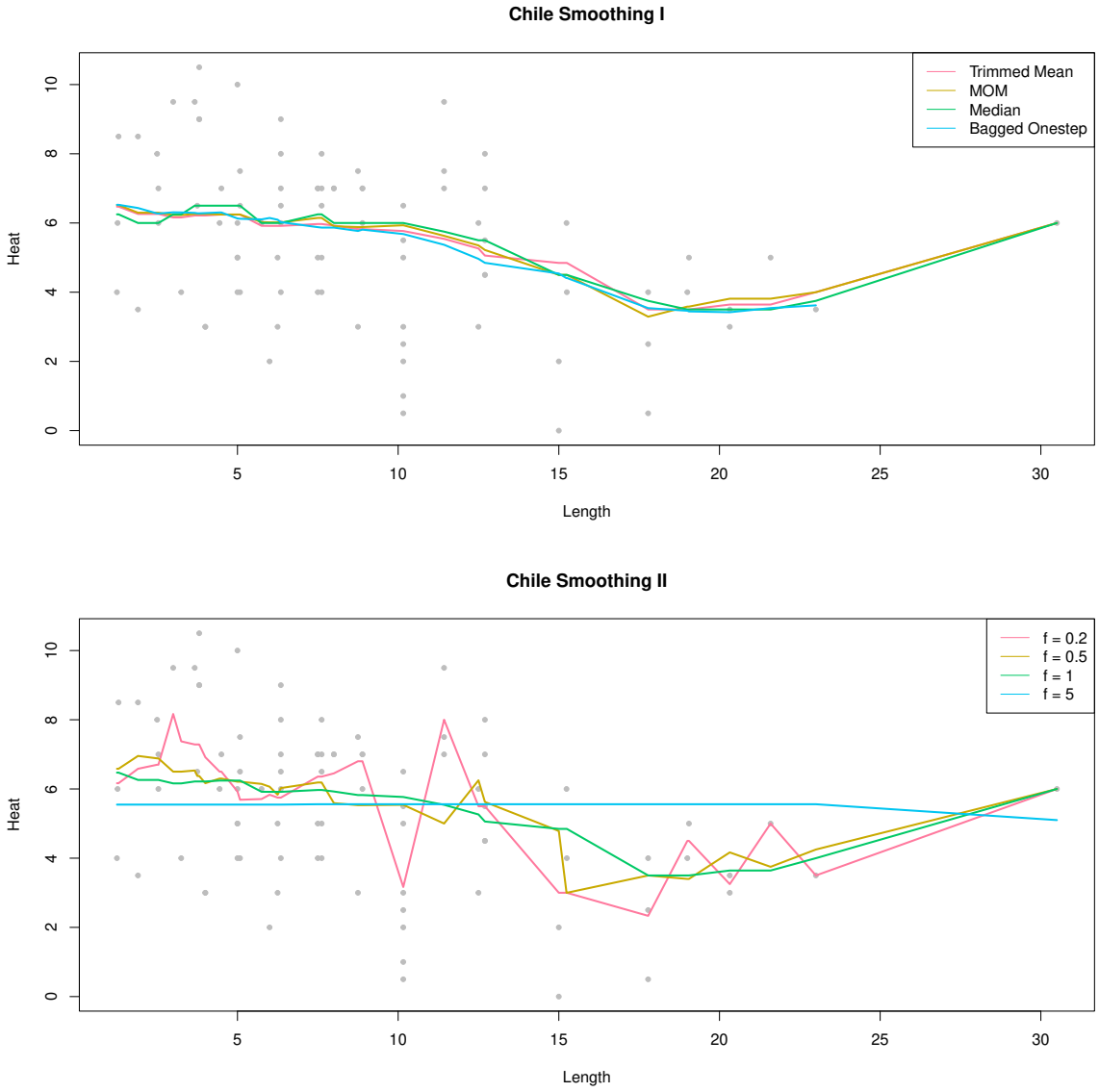

Figure 8: Top panel: smoothers with various robust location measures. Bottom panel: trimmed mean smoother with varying smoothing parameter f.

The WRS2 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 as well (`ancboot`). Both functions perform the running interval smoothing on the trimmed means. Yuen's tests on trimmed mean differences are applied at specified design points. It the design point argument (`pts`) is not specified, the routine automatically computes five points (for details see Wilcox 2017, p. 695). 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 CIs are adjusted to control the probability of at least one Type I error. The p-values are not adjusted.

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 (pre-test and post-test) were recorded. In this analysis we use the pretest score as the covariate and are interested in possible differences between treatment and control group with respect to the post-test scores. We are interested in comparisons at six particular design points. We set the smoothing parameters to a considerably small value.

```
R> comppts <- c(18, 70, 80, 90, 100, 110)
R> fitanc <- ancova(Posttest ~ Pretest + Group, fr1 = 0.3, fr2 = 0.3,
+   data = electric, pts = comppts)
R> fitanc

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

              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 9 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, we get only one significant group difference, for a pre-test 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).

8. Robust mediation analysis

In this section we focus on a simple robust mediator model, involving a response Y, a predictor X, and a mediator M, and consisting of the following set of regressions:

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

The amount of mediation is reflected by the *indirect effect* $\beta_{12}\beta_{23}$ (also called the *mediating effect*). 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).

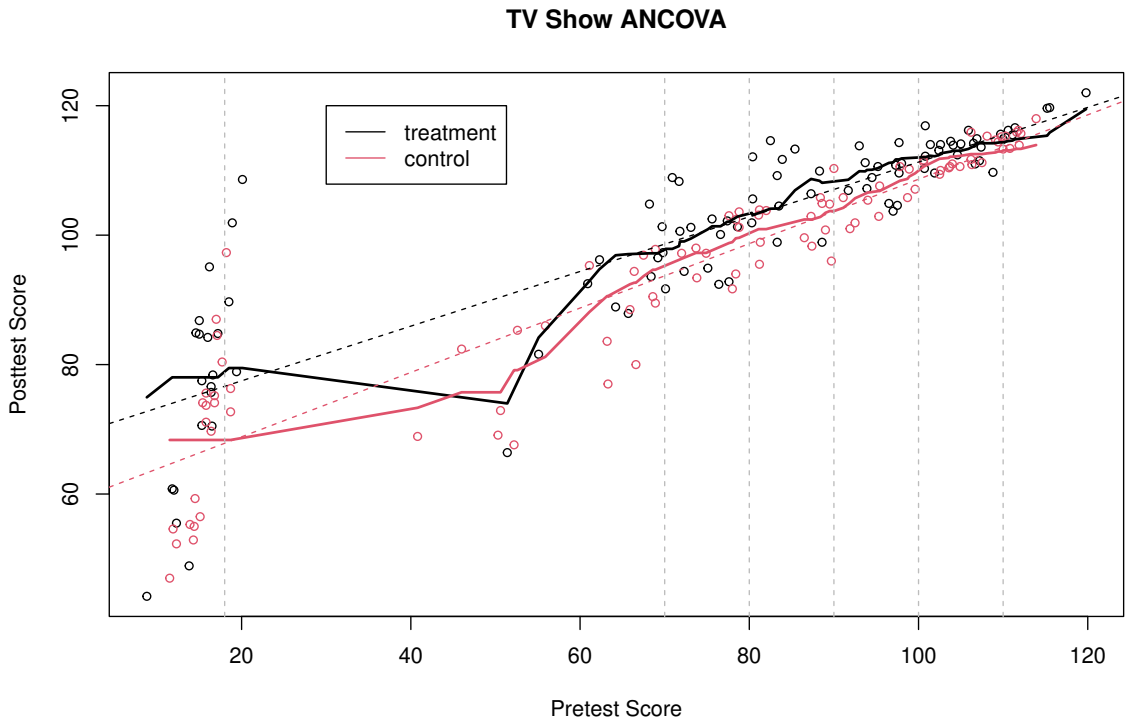

Figure 9: Robust ANCOVA fit on TV show data across treatment and control group. The nonparametric regression lines for both subgroups are shown as well as the OLS fit (dashed lines). The vertical lines show the design points our comparisons are based on.

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 robust approach which is implemented in **WRS2** via the `ZYmediate` function. For technical details we refer to Zu and Yuan (2010).

The example we use for illustration is taken from Howell (2012), and based on data by Leerkes and Crockenberg (2002). In this dataset ($n = 92$), the relationship between how girls were raised by there own mother (`MatCare`) and their later feelings of maternal self-efficacy (`Efficacy`), that is, our belief in our ability to succeed in specific situations, is studied. The mediating variable is self-esteem (`Esteem`). All variables are scored on a continuous scale.

In the first part we fit a standard mediator model with bootstrap-based testing of the mediating effect using the **mediation** package (Tingley, Yamamoto, Hirose, Keele, and Imai 2014).

```
R> library("mediation")
R> fit.mx <- lm(Esteem ~ MatCare, data = Leerkes)
R> fit.yxm <- lm(Efficacy ~ MatCare + Esteem, data = Leerkes)
R> set.seed(123)
R> fitmed <- mediation::mediate(fit.mx, fit.yxm, treat = "MatCare",
```

```
+   mediator = "Esteem", sims = 999, boot = TRUE, boot.ci.type = "bca")
R> summary(fitmed)


Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the BCa Method

               Estimate 95% CI Lower 95% CI Upper p-value
ACME             0.0531       0.0179         0.10   0.006 **
ADE              0.0565      -0.0201         0.13   0.120
Total Effect     0.1096       0.0439         0.18   0.002 **
Prop. Mediated   0.4843       0.2122         1.89   0.008 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 92


Simulations: 999
```

In this output the ACME (average causal mediation effect) represents the indirect effect of `MatCare` on `Efficacy`, including the 95% bootstrap CI. It suggests that there is a significant mediator effect.

Now we fit this mediation model in a robust way with `ZYmediate` from WRS2 which uses bootstrap for the CI of the mediation effect as well.

```
R> set.seed(123)
R> with(Leerkes, ZYmediate(MatCare, Efficacy, Esteem, nboot = 2000))


Call:
ZYmediate(x = MatCare, y = Efficacy, med = Esteem, nboot = 2000)

Mediated effect: 0.0513
Confidence interval: 0.016 0.0979
p-value: 0.001
```

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 basic moderator model can be fitted using `lm`. A robust version of it can be achieved by replacing the `lm` call by an `rlm` call from the **MASS** package.

9. Discussion

This article introduced the **WRS2** package for computing basic robust statistical methods in a user-friendly manner. Such robust models and tests should be used when certain distributional assumptions, as required by classical statistical methods, cannot be justified. The main focus of the **WRS2** package is on simple ANOVA (and related) strategies. For more complex designs, we suggest to consider the following packages. The **robustlmm** package (Koller 2016) implements robust mixed-effects models. For instance, if researchers have to deal with more complex between-within subjects settings that go beyond of what the `bwtrim` function offers, **robustlmm** with its `rlmer` function is highly attractive. For complex mediator-moderator structures, or robust path models with or without latent variables in general, lavaan (Rosseel 2012) offers a variety of robust estimators. Some applications are shown in Field and Wilcox (2017).

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.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Academic Press, New York.

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 AP, Miles J, Field Z (2012). *Discovering Statistics Using R*. Sage Publications, London, UK.

Field AP, Wilcox RR (2017). "Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers." *Behaviour Research and Therapy*, **98**, 19–38.

Gelman A, Hill J (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY.

Harrell FE, Davis CE (1982). "A New Distribution-Free Quantile Estimator." *Biometrika*, **69**, 635–640.

Howell DC (2012). *Statistical Methods for Psychology*. 8th edition. Wadsworth, Belmont, CA.

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

Koller M (2016). "robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models." *Journal of Statistical Software*, **75**(6), 1–24.

Kulinskaya E, Morgenthaler S, Staudte R (2010). "Variance Stabilizing the Difference of Two Binomial Proportions." *The American Statistician*, **64**, 350–356.

Leerkes EM, Crockenberg SC (2002). "The Development of Maternal Self-Efficacy and Its Impact on Maternal Behavior." *Infancy*, **3**, 227–247.

Mair P, Wilcox RR (2020). "Robust Statistical Methods in R Using the WRS2 Package." *Behavior Research Methods*, **52**, 464–488.

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.

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 (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36.

Rousselet GA, Pernet CR, Wilcox RR (2017). "Beyond differences in means: robust graphical methods to compare two groups in neuroscience." *European Journal of Neuroscience*, **46**, 1738–1748.

Seligman MEP, Nolen-Hoeksema S, Thornton N, Thornton CM (1990). "Explanatory Style as a Mechanism of Disappointing Athletic Performance." *Psychological Science*, **1**, 143–146.

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.

Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014). "mediation: R package for causal mediation analysis." *Journal of Statistical Software*, **59**(5), 1–38.

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 (1986). "Improved simultaneous confidence intervals for linear contrasts and regression parameters." *Communications in Statistics - Simulation and Computation*, **15**, 917–932.

Wilcox RR (2017). *Introduction to Robust Estimation & Hypothesis Testing*. 4th edition. Elsevier, Amsterdam, The Netherlands.

Wilcox RR, Erceg-Hurn D (2012). "Comparing two dependent groups via quantiles." *Journal of Applied Statistics*, **39**, 2655–2664.

Wilcox RR, Erceg-Hurn D, Clark F, Carlson M (2014). "Comparing two independent groups via the lower and upper quantiles." *Journal of Statistical Computation and Simulation*, **84**, 1543–1551.

Wilcox RR, Schönbrodt F (2017). *A Package of R. R. Wilcox' Robust Statistics Functions*. R package version 0.34, 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.

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.

Appendix

In this Appendix section we give some technical details on various test statistics using in the text. This part is largely taken from various chapters in Wilcox (2017).

Trimmed/Winsorized mean: Let $W_1, \ldots, W_n$ be the Winsorized random sample based on $X_1, \ldots, X_n$, obtained from replacing the most extreme values (based on Winsorizing level γ) by its neighbors. The *Winsorized mean* is

$$\bar{X}_w = \frac{1}{n}\sum_{i=1}^{n} W_i$$

The *Winsorized variance* is

$$S_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{X}_w)$$

Using this expression, the *standard error of the trimmed mean* can be written as

$$\text{se}(\bar{X}_t) = \frac{S_w}{(1-2\gamma)\sqrt{n}}$$

Yuen's test on trimmed means (`yuen`): Let n_1 and n_2 denote the number of observations in each group, and h_1 and h_2 the number of observations left after trimming. The standard error in the denominator of Eq. (5) is

$$\sqrt{d_1 + d_2} = \sqrt{\frac{(n_1-1)S_{w1}^2}{h_1(h_1-1)} + \frac{(n_2-1)S_{w2}^2}{h_2(h_2-1)}}.$$

The df of the t-distribution the test statistic approximates under the null are

$$\nu_y = \frac{(d_1+d_2)^2}{\frac{d_1^2}{h_1-1} + \frac{d_2^2}{h_2-1}}.$$

The CI is $(\bar{X}_{t1} - \bar{X}_{t2}) \pm t\sqrt{d_1 + d_2}$ where t is the $1 - \alpha/2$ quantile of the t-distribution (with corresponding df).

Robust Cohen's d version (`yuen.effect.ci`): The denominator in the effect size expression in Eq. (6) is

$$S_w^* = \frac{(n_1 - 1)S_{w1}^2 + (n_2 - 1)S_{w2}^2}{n_1 + n_2 - 2}$$

For unequal Winsorized variances Eq. (6) can be replaced by

$$\delta_{t1} = 0.642\frac{\bar{X}_{t1} - \bar{X}_{t2}}{S_{w1}}$$
$$\delta_{t2} = 0.642\frac{\bar{X}_{t1} - \bar{X}_{t2}}{S_{w2}}.$$

Yuen's trimmed means test for dependent samples (`yuend`): Let X_{ij} denote the observed values in group j (here $j = 1, 2$; n observations per group) with trimmed mean $\bar{X}_{tj}$, and Y_{ij} be the Winsorized observations with Winsorized means $\bar{Y}_j$. Let g denote the number of observations Winsorized/trimmed. The effective sample size is $h = n - 2g$. We define the variance term

$$d_j = \frac{1}{h(h-1)}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_j)^2,$$

for groups $j = 1, 2$, and the covariance term

$$d_{12} = \frac{1}{h(h-1)}\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2).$$

The t-distributed test statistic ($df = h - 1$) is given in Eq. (9).

Comparing two discrete distributions (`binband`): The Stoner-Kim method for comparing two distributions (group sizes n_1 and n_2; number of successes r_1 and r_2) defines

$$a_{xy} = \begin{cases} 1 & \text{if } \left|\frac{x}{n_1} - \frac{y}{n_2}\right| \geq \left|\frac{r_1}{n_1} - \frac{r_2}{n_2}\right|, \\ 0 & \text{otherwise.} \end{cases}$$

The test statistic implemented in `binband` is

$$T = \sum_{x=0}^{n_1}\sum_{y=0}^{n_2} a_{xy}B(x; n_1, p)B(y; n_2, p)$$

with $B(\cdot)$ as the probability mass function of the binomial distribution with $p = (r_1 + r_2)/(n_1 + n_2)$. For the CI of the differences in binomial proportions it is referred to Kulinskaya *et al.* (2010).

One-way test trimmed means (`t1way`): For $j = 1, \ldots, J$ groups it uses

$$d_j = \frac{(n_j - 1)S_{wj}^2}{h_j(h_j - 1)}$$

and subsequently computes $w_j = 1/d_j$, $U = \sum_j w_j$, and $\tilde{X} = \frac{1}{U}\sum_j w_j \bar{X}_{tj}$. It follows that

$$A = \frac{1}{J-1}\sum_j w_j \left(\bar{X}_{tj} - \tilde{X}\right)^2,$$
$$B = \frac{2(J-2)}{J^2-1}\sum_j \frac{(1-w_j/U)^2}{h_j - 1}.$$

Based on these components the test statistic as used in `t1way` can be formulated as

$$F_t = \frac{A}{1+B},$$

which is F-distributed with df

$$\nu_1 = J - 1,$$
$$\nu_2 = \left(\frac{3}{J^2-1}\sum_j \frac{(1-w_j/U)^2}{h_j-1}\right)^{-1}.$$

One-way test medians (`med1way`): It follows the same testing strategy as the one for the trimmed means. The starting point is the McKean-Schrader estimate of the squared standard error for the sample median M_j in group j:

$$S_j^2 = \frac{(n_j-1)S_{wj}^2}{h_j - 1}.$$

Subsequently, $w_j = 1/S_j^2$, $U = \sum_j w_j$, and $\tilde{M} = \frac{1}{U}\sum_j w_j M_j$. As above,

$$A = \frac{1}{J-1}\sum_j w_j \left(M_j - \tilde{M}\right)^2,$$
$$B = \frac{2(J-2)}{J^2-1}\sum_j \frac{(1-w_j/U)^2}{n_j - 1}.$$

Based on these components the test statistic as used in `med1way` can be formulated as

$$F_M = \frac{A}{1+B},$$

which, under the null, is F-distributed with df $\nu_1 = J - 1$ and $\nu_2 = \infty$.

Two-way test trimmed means (`t2way`): For a $J \times K$ two-way ANOVA design with factors A and B, let $P = JK$ the total number of cells. The starting point is to construct two contrast matrices, one of dimension $(J-1) \times J$ for factor A, and one of dimension $K-1 \times K$ for factor B. In our 2×3 example we get (see Wilcox 2017, p. 335 for a general construction principle):

$$\mathbf{C}_J = \begin{pmatrix} 1 & -1 \end{pmatrix},$$

and

$$\mathbf{C}_K = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

Now we define two unit vectors of length J and K, i.e., $\mathbf{1}_J$ and $\mathbf{1}_K$. Using these vectors we blow up the contrast matrices using the Kronecker product in order to get a final contrast matrix encoding the main effects for A (dimension $(J-1) \times P$), the main effects for B (dimension $(K-1) \times P$), and the interaction effects (dimension $(K-1) \times p$):

$$\begin{aligned} \mathbf{C}^{(A)} &= \mathbf{C}_J \otimes \mathbf{1}'_K = \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix} \\ \mathbf{C}^{(B)} &= \mathbf{1}'_J \otimes \mathbf{C}_K = \begin{pmatrix} 1 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 1 & -1 \end{pmatrix} \\ \mathbf{C}^{(A\times B)} &= \mathbf{C}_J \otimes \mathbf{C}_K = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix} \end{aligned}$$

In the remainder of this section let $\mathbf{C}$ be a placeholder for either $\mathbf{C}^{(A)}$, $\mathbf{C}^{(B)}$, or $\mathbf{C}^{(A\times B)}$. Let $\mathbf{V}$ be a $P \times P$ diagonal matrix with the squared standard errors of the sample trimmed means on the diagonal. That is,

$$v_{pp} = \frac{(n_p - 1)S^2_{wp}}{h_p(h_p - 1)}.$$

We also define $\bar{\mathbf{X}}'_t = (\bar{X}_{t11}, \bar{X}_{t12}, \ldots, \bar{X}_{t1K}, \bar{X}_{t21}, \bar{X}_{t22}, \ldots, \bar{X}_{t2K}, \ldots, \bar{X}_{tJ1}, \bar{X}_{tJ2}, \ldots, \bar{X}_{tJK})$ as the vector of length p of the sample trimmed means. Based on these matrices we can now define the χ^2-distributed test statistics (main effects for A and B, interaction effect):

$$Q = \bar{\mathbf{X}}'_t \mathbf{C}(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}\mathbf{C}\bar{\mathbf{X}}_t$$

The df's are $J-1$, $K-1$, and $(J-1)(K-1)$, respectively, depending on which effect we study in $\mathbf{C}$. However, the `t2way` function adjusts the critical value c, especially necessary for small sizes. Therefore it does not report any df's. The adjusted critical value is

$$c^* = c + \frac{c}{2k}\left(H\left(1 + \frac{3c}{k+2}\right)\right),$$

where k is the rank of $\mathbf{C}$, $H = \sum_p (r^2_{pp}/(h_p - 1))$, and $\mathbf{R} = \mathbf{V}\mathbf{C}(\mathbf{C}\mathbf{V}\mathbf{C})^{-1}\mathbf{C}$. If $Q \geq c^*$, reject H_0.

Two-way test medians (`med2way`): For the j-th level of factor A and the k-th level of factor B, let n_{jk} be the number of observations, M_{jk} be the sample median with squared standard error S^2_{jk} (McKean-Schrader estimate, see above). We define $R_j = \sum_k M_{jk}$, $W_k = \sum_j M_{jk}$, and $d_{jk} = S^2_{jk}$. We focus on the main effects first. We need

$$\begin{aligned} \hat{\nu}_j &= \frac{\left(\sum_{k=1}^K d_{jk}\right)^2}{\sum_{k=1}^K d^2_{jk}/(n_{jk} - 1)}, \\ \hat{\omega}_k &= \frac{\left(\sum_{j=1}^J d_{jk}\right)^2}{\sum_{j=1}^J d^2_{jk}/(n_{jk} - 1)}. \end{aligned}$$

Let $r_j = 1/\sum_k d_{jk}$ and $w_k = 1/\sum_j d_{jk}$ with sums $r_s = \sum_j r_j$ and $w_s = \sum_k r_k$. Further, $\hat{R} = (\sum_j r_j R_j)/r_s$ and $\hat{W} = (\sum_k w_k W_k)/w_s$. We compute

$$B_a = \sum_{j=1}^{J} \frac{(1 - r_j/r_s)^2}{\hat{\nu}_j}$$
$$B_b = \sum_{k=1}^{K} \frac{(1 - w_j/w_s)^2}{\hat{\omega}_k},$$

which allows us to compute the test statistics for the main effects:

$$V^{(A)} = \frac{\sum_{j=1}^{J} r_j (R_j - \hat{R})^2}{(J-1)\left(1 + \frac{2(J-2)B_a}{J^2-1}\right)}$$
$$V^{(B)} = \frac{\sum_{k=1}^{K} w_k (W_k - \hat{W})^2}{(K-1)\left(1 + \frac{2(K-2)B_b}{K^2-1}\right)}$$

Both statistics are F-distributed with the following df: $\nu_1 = J - 1$ and $\nu_2 = \infty$ for $V^{(A)}$, and $\nu_1 = K - 1$ and $\nu_2 = \infty$ for $V^{(B)}$.

For the $A \times B$ interaction we need $D_{jk} = 1/d_{jk}$, $D_{\cdot k} = \sum_j D_{jk}$, $D_{j\cdot} = \sum_k D_{jk}$, and $D_{\cdot\cdot} = \sum_j \sum_k D_{jk}$. Based on

$$\tilde{M}_{jk} = \sum_{l=1}^{J} D_{lk} M_{lk}/D_{\cdot k} + \sum_{m=1}^{K} D_{jm} M_{jm}/D_{j\cdot} - \sum_{l=1}^{J}\sum_{m=1}^{K} D_{lm} M_{lm}/D_{\cdot\cdot}$$

we define the test statistic

$$V^{(A\times B)} = \sum_{j=1}^{J}\sum_{k=1}^{K} D_{jk}(M_{jk} - \tilde{M}_{jk})^2.$$

This statistic is χ^2-distributed with df $\nu = (J-1)(K-1)$.

One-way repeated measures ANOVA (`rmanova`): Let X_{ij} denote the observed values at time (or group) j with trimmed means $\bar{X}_{tj}$ and $\bar{X}_t = \sum_j \bar{X}_{tj}/J$, and Y_{ij} be the Winsorized observations with Winsorized means $\bar{Y}_{i\cdot}$, $\bar{Y}_{\cdot j}$, and $Y_{\cdot\cdot}$. Let $h = n - 2g$ be the effective sample size based on the trimming amount. We compute

$$Q_c = (n - 2g)\sum_{j=1}^{J}(\bar{X}_{tj} - \bar{X}_t)^2,$$

and

$$Q_e = \sum_{j=1}^{J}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + Y_{\cdot\cdot})^2.$$

Let $R_c = Q_c/(J+1)$ and $R_e = Q_e/((h-1)(J-1))$. The test statistic is

$$F = R_c/R_e.$$

For the df we define

$$v_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{\cdot j})(Y_{ik} - \bar{Y}_{\cdot k}).$$

Let $\bar{v}_{\cdot\cdot} = \sum_j \sum_k v_{jk}/J^2$, $\bar{v}_d = \sum_j v_{jj}/J$, and $\bar{v}_{j\cdot} = \sum_k v_{jk}/J$. Further,

$$A = J^2(\bar{v}_d - \bar{v}_{\cdot\cdot})^2/(J-1),$$
$$B = \sum_{j=1}^{J}\sum_{k=1}^{J} v_{jk}^2 - 2J\sum_{j=1}^{J}\bar{v}_{j\cdot}^2 + J^2\bar{v}_{\cdot\cdot}^2,$$

and

$$\tilde{\epsilon} = \frac{n(J-1)\hat{\epsilon} - 2}{(J-1)(n-1-(J-1)\hat{\epsilon})}$$

with $\hat{\epsilon} = A/B$. Subsequently, the df can be expressed as

$$\nu_1 = (J-1)\tilde{\epsilon},$$
$$\nu_2 = (J-1)(h-1)\tilde{\epsilon}.$$

Between-within subjects ANOVA on the trimmed means (`bwtrim`): The test statistic is constructed according to the same principles as in `t2way`. The main difference is that for each factor level j of factor A we estimate

$$\mathbf{V}_j = \frac{(n_j - 1)\mathbf{S}_j}{h_j(h_j - 1)},$$

where $\mathbf{S}_j$ is an estimate for the $K \times K$ Winsorized covariance matrix. The $\mathbf{V}_j$ matrices are collected in the block diagonal matrix $\mathbf{V}$. Let $\mathbf{C}$ be the contrast matrix (rank k) of the effect we want to study. The test statistic is

$$Q = \bar{\mathbf{X}}_t'\mathbf{C}(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}\mathbf{C}\bar{\mathbf{X}}_t.$$

This statistic needs to be modified as follows in order to be F-distributed. Let $\mathbf{Q}_j$ be a $JK \times JK$ a block diagonal matrix. We compute

$$A = \frac{1}{2}\sum_{j=1}^{J}(\mathrm{tr}((\mathbf{V}\mathbf{C}'(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}\mathbf{C}\mathbf{Q}_j)^2) + (\mathrm{tr}(\mathbf{V}\mathbf{C}'(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}\mathbf{C}\mathbf{Q}_j))^2)/(h_j - 1),$$

and

$$c = k + 2A - \frac{6A}{k+2}.$$

Under H_0, Q/c is F-distributed with df $\nu_1 = k$ and $\nu_2 = k(k+2)/(3A)$.

Hochberg's method for controlling the FWE: Let $p_{[1]}, \ldots, p_{[C]}$ be the p-values associated with C tests, in descending order. Let α be the significance level. The procedure starts with rejecting all hypotheses if $p_{[k]} \leq \alpha/k$ for $k = 1$. If $p_{[1]} > \alpha$, set $k := k + 1$. Again, apply $p_{[k]} \leq \alpha/k$. If $p_{[k]} > \alpha/k$, increment k. Repeat until either all hypotheses under consideration are rejected or all C hypotheses have been tested.

Affiliation:

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