The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)

Thorsten Pohlert

latest revision: 2015-12-07

Contents

1 Introduction

The Kruskal and Wallis one-way analysis of variance by ranks or van der Waerden's normal score test can be employed, if the data do not meet the assumptions for one-way ANOVA. Provided that significant differences were detected by the omnibus test, one may be interested in applying post-hoc tests for pairwise multiple comparisons (such as Nemenyi's test, Dunn's test, Conover's test, van der Waerden's test). Similarly, one-way ANOVA with repeated measures that is also referred to as ANOVA with unreplicated block design can also be conducted via the Friedman-Test or the Quade-test. The consequent post-hoc pairwise multiple comparison tests according to Nemenyi, Conover and Quade are also provided in this package. Finally Durbin's test for a two-way balanced incomplete block design (BIBD) is also given in this package.

2 Comparison of multiple independent samples (One-factorial design)

2.1 Kruskal and Wallis test

The linear model of a one-way layout can be written as:

$$y_i = \mu + \alpha_i + \epsilon_i, \tag{1}$$

with y the response vector, μ the global mean of the data, α_i the difference to the mean of the i-th group and ϵ the residual error. The non-parametric alternative is the Kruskal

and Wallis test. It tests the null hypothesis, that each of the k samples belong to the same population ($H_0 : \bar{R}_{i.} = (n+1)/2$). First, the response vector y is transformed into ranks with increasing order. In the presence of sequences with equal values (i.e. ties), mean ranks are designated to the corresponding realizations. Then, the test statistic can be calculated according to Eq. 2:

$$\hat{H} = \left[\frac{12}{n\,(n+1)}\right]\left[\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3\,(n+1) \tag{2}$$

with $n = \sum_i^k n_i$ the total sample size, n_i the number of data of the i-th group and R_i^2 the squared rank sum of the i-th group. In the presence of many ties, the test statistics $\hat{H}$ can be corrected using Eqs. 3 and 4

$$C = 1 - \frac{\sum_{i=1}^{i=r}\left(t_i^3 - t_i\right)}{n^3 - n}, \tag{3}$$

with t_i the number of ties of the i-th group of ties.

$$\hat{H}^* = \hat{H}/C \tag{4}$$

The Kruskal and Wallis test can be employed as a global test. As the test statistic $\bar{H}$ is approximately χ^2-distributed, the null hypothesis is withdrawn, if $\hat{H} > \chi^2_{k-1;\alpha}$. It should be noted, that the tie correction has only a small impact on the calculated statistic and its consequent estimation of levels of significance.

2.2 Kruskal-Wallis – post-hoc tests after Nemenyi

Provided, that the globally conducted Kruskal-Wallis test indicates significance (i.e. H_0 is rejected, and H_A : 'at least on of the k samples does not belong to the same population' is accepted), one may be interested in identifying which group or groups are significantly different. The number of pairwise contrasts or subsequent tests that need to be conducted is $m = k\,(k-1)\,/2$ to detect the differences between each group. Nemenyi proposed a test that originally based on rank sums and the application of the *family-wise error* method to control Type I error inflation, if multiple comparisons are done. The Tukey and Kramer approach uses mean rank sums and can be employed for equally as well as unequally sized samples without ties (Sachs, 1997, p. 397). The null hypothesis $H_0 : \bar{R}_i = \bar{R}_j$ is rejected, if a critical absolute difference of mean rank sums is exceeded.

$$|\bar{R}_i - \bar{R}_j| > \frac{q_{\infty;k;\alpha}}{\sqrt{2}}\sqrt{\left[\frac{n\,(n+1)}{12}\right]\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}, \tag{5}$$

where $q_{\infty;k;\alpha}$ denotes the upper quantile of the studentized range distribution. Although these quantiles can not be computed analytically, as $df = \infty$, a good approximation is to set df very large: such as $q_{1000000;k;\alpha} \sim q_{\infty;k;\alpha}$. This inequality (5) leads to the same critical differences of rank sums ($|R_i - R_j|$) when multiplied with n for

$\alpha = [0.1, 0.5, 0.01]$, as reported in the tables of Wilcoxon and Wilcox (1964, pp. 29–31). In the presence of ties the approach presented by Sachs (1997, p. 395) can be employed (6), provided that $(n_i, n_j, \ldots, n_k \geq 6)$ and $k \geq 4$.:

$$|\bar{R}_i - \bar{R}_j| > \sqrt{C\chi^2_{k-1;\alpha}\left[\frac{n\,(n+1)}{12}\right]\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}, \tag{6}$$

where C is given by Eq. 3. The function `posthoc.kruskal.nemenyi.test` does not evaluate the critical differences as given by Eqs. 5 and 6, but calculates the corresponding level of significance for the estimated statistics q and χ^2, respectively.

In the special case, that several treatments shall only be tested against one control experiment, the number of tests reduces to $m = k - 1$. This case is given in section 2.8.

2.3 Examples using `posthoc.kruskal.nemenyi.test`

The function `kruskal.test` is provided with the library `stats` (R Core Team, 2013). The data-set `InsectSprays` was derived from a one factorial experimental design and can be used for demonstration purposes. Prior to the test, a visualization of the data (Fig 1) might be helpful:

Based on a visual inspection, one can assume that the insecticides *A, B, F* differ from *C, D, E*. The global test can be conducted in this way:

```
> kruskal.test(count ~ spray, data=InsectSprays)

        Kruskal-Wallis rank sum test

data:  count by spray
Kruskal-Wallis chi-squared = 54.6913, df = 5, p-value = 1.511e-10
```

As the Kruskal-Wallis Test statistics is highly significant ($\chi^2(5) = 54.69, p < 0.01$), the null hypothesis is rejected. Thus, it is meaningful to apply post-hoc tests with the function `posthoc.kruskal.nemenyi.test`.

```
> require(PMCMR)
> data(InsectSprays)
> attach(InsectSprays)
> posthoc.kruskal.nemenyi.test(x=count, g=spray, dist="Tukey")

        Pairwise comparisons using Tukey and Kramer (Nemenyi) test
                    with Tukey-Dist approximation for independent samples

data:  count and spray

  A       B       C       D       E
B 0.99961 -       -       -       -
```

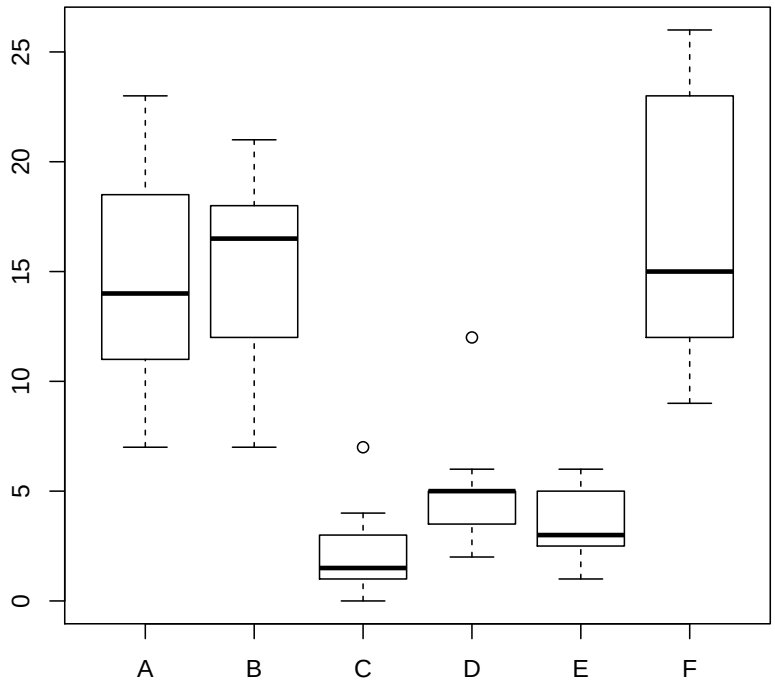

Figure 1: Boxplot of the `InsectSprays` data set.

```
C 2.8e-05 5.7e-06 -       -       -
D 0.02293 0.00813 0.56300 -       -
E 0.00169 0.00047 0.94109 0.97809 -
F 0.99861 1.00000 3.5e-06 0.00585 0.00031

P value adjustment method: none
```

The test returns the lower triangle of the matrix that contains the p-values of the pairwise comparisons. Thus $|\bar{R}_A - \bar{R}_B| : n.s.$, but $|\bar{R}_A - \bar{R}_C| : p < 0.01$. Since PMCMR-1.1 there is a formula method included. Thus the test can also be conducted in the following way:

```
> posthoc.kruskal.nemenyi.test(count ~ spray, data=InsectSprays, dist="Tukey")

        Pairwise comparisons using Tukey and Kramer (Nemenyi) test
                   with Tukey-Dist approximation for independent samples
```

```
data:  count by spray

  A       B       C       D       E
B 0.99961 -       -       -       -
C 2.8e-05 5.7e-06 -       -       -
D 0.02293 0.00813 0.56300 -       -
E 0.00169 0.00047 0.94109 0.97809 -
F 0.99861 1.00000 3.5e-06 0.00585 0.00031

P value adjustment method: none
```

As there are ties present in the data, one may also conduct the Chi-square approach:

```
> (out <- posthoc.kruskal.nemenyi.test(x=count, g=spray, dist="Chisquare"))

        Pairwise comparisons using Nemenyi-test with Chi-squared
                       approximation for independent samples

data:  count and spray

  A       B       C       D       E
B 0.99985 -       -       -       -
C 0.00037 9.4e-05 -       -       -
D 0.08359 0.03812 0.73938 -       -
E 0.01113 0.00391 0.97354 0.99070 -
F 0.99945 1.00000 6.2e-05 0.02955 0.00281

P value adjustment method: none
```

which leads to different levels of significance, but to the same test decision. The arguments of the returned object of class `pairwise.h.test` can be further explored. The statistics, in this case the χ^2 estimations, can be taken in this way:

```
> print(out$statistic)

            A            B          C          D        E
B  0.09741248           NA         NA         NA       NA
C 22.70093702 25.772474315         NA         NA       NA
D  9.68046043 11.720034247  2.7330908         NA       NA
E 14.76750381 17.263698630  0.8495291  0.5351027       NA
F  0.16383657  0.008585426 26.7218417 12.3630375 18.04226
```

2.4 Kruskal-Wallis – post-hoc test after Dunn

Dunn (1964) has proposed a test for multiple comparisons of rank sums based on the z-statistics of the standard normal distribution. The null hypothesis ($H0$), the probability

of observing a randomly selected value from the first group that is larger than a randomly selected value from the second group equals one half, is rejected, if a critical absolute difference of mean rank sums is exceeded:

$$|\bar{R}_i - \bar{R}_j| > z_{1-\alpha/2*} \sqrt{\left[\frac{n\,(n+1)}{12} - B\right]\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}, \tag{7}$$

with $z_{1-\alpha/2*}$ the value of the standard normal distribution for a given adjusted $\alpha/2*$ level depending on the number of tests conducted and B the correction term for ties, which was taken from Glantz (2012) and is given by Eq. 8:

$$B = \frac{\sum_{i=1}^{i=r}\left(t_i^3 - t_i\right)}{12\,(n-1)} \tag{8}$$

The function `posthoc.kruskal.dunn.test` does not evaluate the critical differences as given by Eqs. 7, but calculates the corresponding level of significance for the estimated statistics z, as adjusted by any method implemented in `p.adjust` to account for Type I error inflation. It should be noted that Dunn (1964) originally used a Bonferroni adjustment of p-values. For this specific case, the test is sometimes referred as to the Bonferroni-Dunn test.

2.5 Example using `posthoc.kruskal.dunn.test`

We can go back to the example with InsectSprays.

```
> require(PMCMR)
> data(InsectSprays)
> attach(InsectSprays)
> posthoc.kruskal.dunn.test(x=count, g=spray, p.adjust.method="none")

        Pairwise comparisons using Dunn's-test for multiple
                        comparisons of independent samples

data:  count and spray

  A       B       C       D       E
B 0.75448 -       -       -       -
C 1.8e-06 3.6e-07 -       -       -
D 0.00182 0.00060 0.09762 -       -
E 0.00012 3.1e-05 0.35572 0.46358 -
F 0.68505 0.92603 2.2e-07 0.00043 2.1e-05

P value adjustment method: none
```

The test returns the lower triangle of the matrix that contains the p-values of the pairwise comparisons. As `p.adjust.method="none"`, the p-values are not adjusted.

Hence, there is a Type I error inflation that leads to a false positive discovery rate. This can be solved by applying e.g. a Bonferroni-type adjustment of p-values.

```
> require(PMCMR)
> data(InsectSprays)
> attach(InsectSprays)
> posthoc.kruskal.dunn.test(x=count, g=spray, p.adjust.method="bonferroni")

        Pairwise comparisons using Dunn's-test for multiple
                        comparisons of independent samples

data:  count and spray

  A       B       C       D       E
B 1.00000 -       -       -       -
C 2.7e-05 5.5e-06 -       -       -
D 0.02735 0.00904 1.00000 -       -
E 0.00177 0.00047 1.00000 1.00000 -
F 1.00000 1.00000 3.3e-06 0.00640 0.00031

P value adjustment method: bonferroni
```

2.6 Kruskal-Wallis – post-hoc test after Conover

Conover and Iman (1979) proposed a test that aims at having a higher test power than the tests given by inequalities 5 and 6:

$$|\bar{R}_i - \bar{R}_j| > t_{1-\alpha/2;n-k} \sqrt{s^2 \left[\frac{n-1-\hat{H}^*}{n-k}\right] \left[\frac{1}{n_i} + \frac{1}{n_j}\right]}, \tag{9}$$

with $\hat{H}^*$ the tie corrected Kruskal-Wallis statistic according to Eq. 4 and $t_{1-\alpha/2;n-k}$ the t-value of the student-t-distribution. The variance s^2 is given in case of ties by:

$$s^2 = \frac{1}{n-1} \left[\sum R_i^2 - n\frac{(n+1)^2}{4}\right] \tag{10}$$

The variance s^2 simplifies to $n\,(n+1)\,/12$, if there are no ties present. Although Conover and Iman (1979) did not propose an adjustment of p-values, the function `posthoc.kruskal.conover.test` has implemented methods for p-adjustment from the function `p.adjust`.

2.7 Example using `posthoc.kruskal.conover.test`

```
> require(PMCMR)
> data(InsectSprays)
```

```
> attach(InsectSprays)
> posthoc.kruskal.conover.test(x=count, g=spray, p.adjust.method="none")

        Pairwise comparisons using Conover's-test for multiple
                        comparisons of independent samples

data:  count and spray

  A       B       C       D       E
B 0.5314  -       -       -       -
C 3.7e-14 3.0e-15 -       -       -
D 3.1e-08 2.4e-09 0.0014  -       -
E 7.5e-11 5.6e-12 0.0676  0.1451  -
F 0.4175  0.8524  1.4e-15 1.1e-09 2.6e-12

P value adjustment method: none

> require(PMCMR)
> data(InsectSprays)
> attach(InsectSprays)
> posthoc.kruskal.conover.test(x=count, g=spray, p.adjust.method="bonferroni")

        Pairwise comparisons using Conover's-test for multiple
                        comparisons of independent samples

data:  count and spray

  A       B       C       D       E
B 1.000   -       -       -       -
C 5.6e-13 4.5e-14 -       -       -
D 4.7e-07 3.6e-08 0.021   -       -
E 1.1e-09 8.5e-11 1.000   1.000   -
F 1.000   1.000   2.1e-14 1.7e-08 3.9e-11

P value adjustment method: bonferroni
```

2.8 Dunn's multiple comparison test with one control

Dunn's test (see section 2.4), can also be applied for multiple comparisons with one control (Siegel and Castellan Jr., 1988):

$$|\bar{R}_0 - \bar{R}_j| > z_{1-\alpha/2*} \sqrt{\left[\frac{n\,(n+1)}{12} - B\right]\left[\frac{1}{n_0} + \frac{1}{n_j}\right]}, \tag{11}$$

where $\bar{R}_0$ denotes the mean rank sum of the control experiment. In this case the number of tests is reduced to $m = k - 1$, which changes the p-adjustment according to Bonferroni (or others). The function `dunn.test.control` employs this test, but **the user need to be sure that the control is given as the first level in the group vector**.

2.9 Example using `dunn.test.control`

We can use the `PlantGrowth` dataset, that comprises data with dry matter weight of yields with one control experiment (i.e. no treatment) and to different treatments. In this case we are only interested, whether the treatments differ significantly from the control experiment.

```
> require(stats)
> data(PlantGrowth)
> attach(PlantGrowth)
> kruskal.test(weight, group)

        Kruskal-Wallis rank sum test

data:  weight and group
Kruskal-Wallis chi-squared = 7.9882, df = 2, p-value = 0.01842

> dunn.test.control(x=weight,g=group, p.adjust="bonferroni")

        Pairwise comparisons using Dunn's-test for multiple
                        comparisons with one control

data:  weight and group

     ctrl
trt1 0.53
trt2 0.18

P value adjustment method: bonferroni
```

According to the global Kruskal-Wallis test, there are significant differences between the groups, $\chi^2(2) = 7.99, p < 0.05$. However, the Dunn-test with Bonferroni adjustment of p-values shows, that there are no significant effects.

If one may cross-check the findings with ANOVA and multiple comparison with one control using the LSD-test, he/she can do the following:

```
> summary.lm(aov(weight ~ group))

Call:
aov(formula = weight ~ group)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.1971  25.527   <2e-16 ***
grouptrt1    -0.3710     0.2788  -1.331   0.1944
grouptrt2     0.4940     0.2788   1.772   0.0877 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.2641,        Adjusted R-squared:  0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

The last line provides the statistics for the global test, i.e. there is a significant treatment effect according to one-way ANOVA, $F(2, 27) = 4.85, p < 0.05, \eta^2 = 0.264$. The row that starts with `Intercept` gives the group mean of the control, its standard error, the t-value for testing $H0 : \mu = 0$ and the corresponding level of significance. The following lines provide the difference between the averages of the treatment groups with the control, where $H0 : \mu_0 - \mu_j = 0$. Thus the `trt1` does not differ significantly from the `ctr`, $t = -1.331, p = 0.194$. There is a significant difference between `trt2` and `ctr` as indicated by $t = 1.772, p < 0.1$.

2.10 van der Waerden test

The van der Waerden test can be used as an alternative to the Kruskal-Wallis test, if the data to not meet the requirements for ANOVA (Conover and Iman, 1979). Let the Kruskal-Wallis ranked data denote $R_{i,j}$, then the normal scores $A_{i,j}$ are derived from the standard normal distribution according to Eq. 12.

$$A_{i,j} = \phi^{-1}\left(\frac{R_{i,j}}{n+1}\right) \tag{12}$$

Let the sum of the i-th score denote A_j. The variance S^2 is calculated as given in Eq. 13.

$$S^2 = \frac{1}{n-1}\sum A_{i,j}^2 \tag{13}$$

Finally the test statistic is given by Eq. 14.

$$T = \frac{1}{S^2}\sum_{j=1}^{k}\frac{A_j^2}{n_j} \tag{14}$$

The test statistic T is approximately χ^2-distributed and tested for significance on a level of $1-\alpha$ with $df = k-1$.

2.11 Example using `vanWaerden.test`

```
> require(PMCMR)
> data(InsectSprays)
> attach(InsectSprays)
> vanWaerden.test(x=count, g=spray)

        Van der Waerden normal scores test

data:  count and spray
Van der Waerden chi-squared = 50.302, df = 5, p-value = 1.202e-09
```

2.12 post-hoc test after van der Waerden for multiple pairwise comparisons

Provided that the global test according to van der Waerden indicates significance, multiple comparisons can be done according to the inequality 15.

$$\|\frac{A_i}{n_i} - \frac{A_j}{n_j}\| > t_{1-\alpha/2*;n-k}\sqrt{S^2 \ \frac{n-1-T}{n-k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{15}$$

The test given in Conover and Iman (1979) does not adjust p-values. However, the function has included the methods for p-value adjustment as given by `p.adjust`.

2.13 Example using `posthoc.vanWaerden.test`

```
> require(PMCMR)
> data(InsectSprays)
> attach(InsectSprays)
> posthoc.vanWaerden.test(x=count, g=spray, p.adjust.method="none")

        Pairwise comparisons using van der Waerden normal scores test for
              multiple comparisons of independent samples

data:  count and spray

  A       B       C       D       E
B 0.6366  -       -       -       -
C 6.9e-12 9.8e-13 -       -       -
D 9.0e-06 1.5e-06 0.0008  -       -
E 5.5e-08 8.1e-09 0.0316  0.1919  -
F 0.2323  0.4675  5.0e-14 8.6e-08 4.1e-10

P value adjustment method: none
```

3 Test of k independent samples against an ordered alternative

3.1 Jonckheere-Terpstrata test

For testing k independent samples against an ordered alternative (e.g. k groups with increasing doses of treatment), the test according to Jonckheere (1954) can be employed. Both, the null and alternative hypothesis can be formulated as population medians (θ_i) for k populations ($k > 2$). Thus the null hypothesis, $H0 : \theta_1 = \theta_2 = \ldots = \theta_k$ is tested e.g. in the one-sided case against the alternative, $HA : \theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$, with at least one strict inequality.

The statistic J is calculated according to Eq. 16.

$$J = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij} \tag{16}$$

where U_{ij} is defined as:

$$U_{ij} = \sum_{s=1}^{n_i} \sum_{t=1}^{n_j} \psi \left(X_{jt} - X_{is} \right) \tag{17}$$

with

$$\psi \left(u \right) = \begin{cases} 1 & \forall \quad u > 0 \\ 1/2 & \forall \quad u = 0 \\ 0 & \forall \quad u < 0 \end{cases} \tag{18}$$

The mean μ_J is given by

$$\mu_J = \frac{N^2 - \sum_{i=1}^{k} n_i^2}{4} \tag{19}$$

and the variance σ_J is defined as

$$\sigma_J = \sqrt{\frac{N^2 \left(2N + 3 \right) - \sum_{i=1}^{k} n_i^2 \left(2n_i + 3 \right)}{72}}. \tag{20}$$

The $\hat{z}$-value of the standard normal distribution is calculated as:

$$\hat{z} = \frac{J - \mu_J}{\sigma_J} \tag{21}$$

For a one-sided test, the $H0$ is rejected, if $\hat{z} > z_{1-\alpha/2}$. The implemented function can test for monotonic trend (two-sided), increasing or decreasing trend (one-sided).

3.2 Example using `jonckheere.test`

```
> ## Example from Sachs (1997, p. 402)
> require(PMCMR)
> x <- c(106, 114, 116, 127, 145, 110, 125,
+        143, 148, 151, 136, 139, 149, 160,
+        174)
> g <- as.factor(c(rep(1,5), rep(2,5), rep(3,5)))
> levels(g) <- c("A", "B", "C")
> jonckheere.test(x , g, "increasing")

        Jonckheere-Terpstrata test

data:  x and g
Jonckheere z-value = 2.2716, p-value = 0.01156
alternative hypothesis: increasing

> rm(x,g)
```

4 Comparison of multiple joint samples (Two-factorial unreplicated complete block design)

4.1 Friedman test

The linear model of a two factorial unreplicated complete block design can be written as:

$$y_{i,j} = \mu + \alpha_i + \pi_j + \epsilon_{i,j} \tag{22}$$

with π_j the j-th level of the block (e.g. the specific response of the j-th test person). The Friedman test is the non-parametric alternative for this type of k dependent treatment groups with equal sample sizes. The null hypothesis, $H0 : F(1) = F(2) = \ldots = F(k)$ is tested against the alternative hypothesis: at least one group does not belong to the same population. The response vector y has to be ranked in ascending order separately for each block $\pi_j : j = 1, \ldots m$. After that, the statistics of the Friedman test is calculated according to Eq. 23:

$$\hat{\chi}^2_R = \left[\frac{12}{nk\,(k+1)} \sum_{i=1}^{k} R_i \right] - 3n\,(k+1) \tag{23}$$

The Friedman statistic is approximately χ^2-distributed and the null hypothesis is rejected, if $\hat{\chi}_R > \chi^2_{k-1;\alpha}$.

4.2 Friedman – post-hoc test after Nemenyi

Provided that the Friedman test indicates significance, the post-hoc test according to Nemenyi (1963) can be employed (Sachs, 1997, p. 668). This test requires a balanced design ($n_1 = n_2 = \ldots = n_k = n$) for each group k and a Friedman-type ranking of the data. The inequality 24 was taken from Demsar (2006, p. 11), where the critical difference refer to mean rank sums ($|\bar{R}_i - \bar{R}_j|$):

$$|\bar{R}_i - \bar{R}_j| > \frac{q_{\infty;k;\alpha}}{\sqrt{2}} \sqrt{\frac{k\,(k+1)}{6n}} \tag{24}$$

This inequality leads to the same critical differences of rank sums ($|R_i - R_j|$) when multiplied with n for $\alpha = [0.1, 0.5, 0.01]$, as reported in the tables of Wilcoxon and Wilcox (1964, pp. 36–38). Likewise to the `posthoc.kruskal.nemenyi.test` the function `posthoc.friedman.nemenyi.test` calculates the corresponding levels of significance and the generic function `print` depicts the lower triangle of the matrix that contains these p-values. The test according to Nemenyi (1963) was developed to account for a family-wise error and is already a conservative test. This is the reason, why there is no p-adjustment included in the function.

4.3 Example using `posthoc.friedman.nemenyi.test`

This example is taken from Sachs (1997, p. 675) and is also included in the help page of the function `posthoc.friedman.nemenyi.test`. In this experiment, six persons (block) subsequently received six different diuretics (groups) that are denoted A to F. The responses are the concentration of Na in urine measured two hours after each treatment.

```
> require(PMCMR)
> y <- matrix(c(
+ 3.88, 5.64, 5.76, 4.25, 5.91, 4.33, 30.58, 30.14, 16.92,
+ 23.19, 26.74, 10.91, 25.24, 33.52, 25.45, 18.85, 20.45,
+ 26.67, 4.44, 7.94, 4.04, 4.4, 4.23, 4.36, 29.41, 30.72,
+ 32.92, 28.23, 23.35, 12, 38.87, 33.12, 39.15, 28.06, 38.23,
+ 26.65),nrow=6, ncol=6,
+ dimnames=list(1:6,c("A","B","C","D","E","F")))
> print(y)

     A     B     C    D     E     F
1 3.88 30.58 25.24 4.44 29.41 38.87
2 5.64 30.14 33.52 7.94 30.72 33.12
3 5.76 16.92 25.45 4.04 32.92 39.15
4 4.25 23.19 18.85 4.40 28.23 28.06
5 5.91 26.74 20.45 4.23 23.35 38.23
6 4.33 10.91 26.67 4.36 12.00 26.65
```

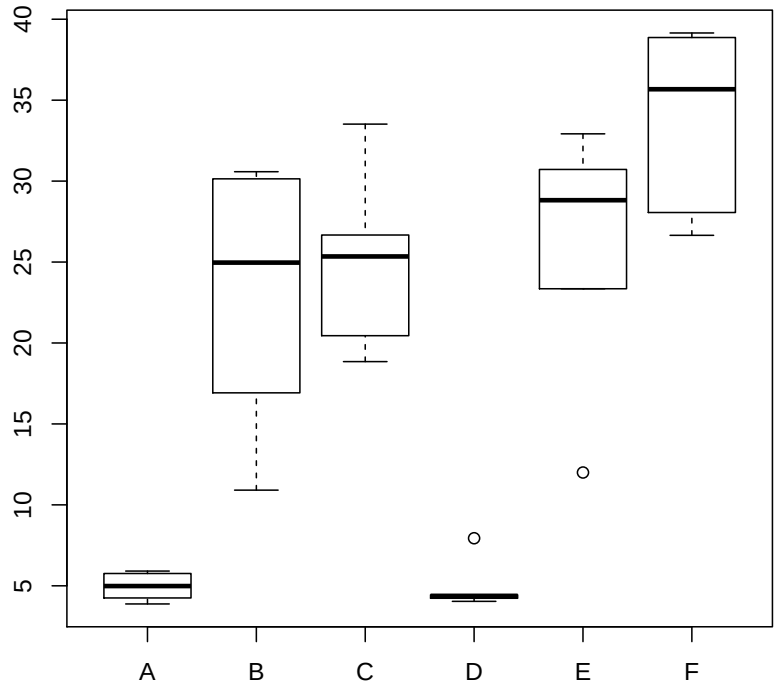

Figure 2: Na-concentration (mval) in urine of six test persons after treatment with six different diuretics.

Based on a visual inspection (Fig. 2), one may assume different responses of Na-concentration in urine as related to the applied diuretics.

The global test is the Friedman test, that is already implemented in the package `stats` (R Core Team, 2013):

```
> friedman.test(y)

        Friedman rank sum test

data:  y
Friedman chi-squared = 23.3333, df = 5, p-value = 0.0002915
```

As the Friedman test indicates significance ($\chi^2(5) = 23.3, p < 0.01$), it is meaningful to conduct multiple comparisons in order to identify differences between the diuretics.

```
> posthoc.friedman.nemenyi.test(y)

        Pairwise comparisons using Nemenyi multiple comparison test
             with q approximation for unreplicated blocked data

data:  y

  A      B      C      D      E
B 0.1880 -      -      -      -
C 0.0917 0.9996 -      -      -
D 0.9996 0.3388 0.1880 -      -
E 0.0395 0.9898 0.9996 0.0917 -
F 0.0016 0.6363 0.8200 0.0052 0.9400

P value adjustment method: none
```

According to the Nemenyi post-hoc test for multiple joint samples, the treatment F based on the Na diuresis differs highly significant ($p < 0.01$) to A and D, and E differs significantly ($p < 0.05$) to A. Other contrasts are not significant ($p > 0.05$). This is the same test decision as given by (Sachs, 1997, p. 675).

4.4 Friedman – post-hoc test after Conover

Conover (1999) proposed a post-hoc test for pairwise comparisons, if Friedman-Test indicated significance. The absolute difference between two group rank sums are significantly different, if the following inequality is satisfied:

$$|R_i - R_j| > t_{1-\alpha/2;(n-1)(k-1)} \sqrt{\frac{2k\left(1 - \frac{\hat{\chi}^2_R}{n(k-1)}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{k} R_{i,j}^2 - \frac{nk(k+1)^2}{4}\right)}{(k-1)(n-1)}} \tag{25}$$

Although Conover (1999) originally did not include a p-adjustment, the function has included the methods as given by `p.adjust`, because it is a very liberal test. So it is up to the user, to apply a p-adjustment or not, when using this function.

4.5 Example using `posthoc.friedman.conover.test`

```
> require(PMCMR)
> y <- matrix(c(
+ 3.88, 5.64, 5.76, 4.25, 5.91, 4.33, 30.58, 30.14, 16.92,
+ 23.19, 26.74, 10.91, 25.24, 33.52, 25.45, 18.85, 20.45,
+ 26.67, 4.44, 7.94, 4.04, 4.4, 4.23, 4.36, 29.41, 30.72,
+ 32.92, 28.23, 23.35, 12, 38.87, 33.12, 39.15, 28.06, 38.23,
+ 26.65),nrow=6, ncol=6,
```

```
+ dimnames=list(1:6,c("A","B","C","D","E","F")))
> friedman.test(y)

        Friedman rank sum test

data:  y
Friedman chi-squared = 23.3333, df = 5, p-value = 0.0002915

> posthoc.friedman.conover.test(y=y, p.adjust="none")

        Pairwise comparisons using Conover's test for a two-way
                    balanced complete block design

data:  y

  A       B       C       D       E
B 0.00014 -       -       -       -
C 3.0e-05 0.55547 -       -       -
D 0.55547 0.00067 0.00014 -       -
E 6.5e-06 0.24321 0.55547 3.0e-05 -
F 8.0e-08 0.00621 0.02468 3.3e-07 0.08511

P value adjustment method: none
```

4.6 Quade test

Likewise to the Friedman test, the Quade test is a non-parametric test for analyzing randomized complete block designs. According to Conover (1999) the Quade test is more powerful than the Friedman test in the case of $k < 5$. The $H0$ is that the k groups are identical, whereas HA means that at least one group differs from at least one other group. First, the data are ranked within each block (i.e. row) to yield $R_{i,j}$. Then, the range in each block (maximum minus minimum value of the data) needs to be computed and ranked, Q_i. Then the scores are

$$S_{i,j} = Q_i * (R_{i,j} - (k+1)/2) \tag{26}$$

and

$$S_j = \sum_{i=1}^{b} S_{i,j} \tag{27}$$

Then the test statistic is

$$\hat{F} = \frac{(b-1)\,B}{A-B} \tag{28}$$

with

$$\begin{aligned} A &= \sum_{i=1}^{b}\sum_{j=1}^{k} S_{i,j}^2 \\ B &= \frac{1}{b}\sum_{i=1}^{k} S_j^2 \end{aligned}$$

The test statistic $\hat{F}$ is tested against the F-quantile for a given α with $df_1 = k - 1$ and $df_2 = (b-1)(k-1)$. This is equivalent to a two-way ANOVA on the scores $S_{i,j}$.

4.7 Quade – posthoc tests

The package `PMCMR` offers two posthoc tests following a significant Quade test. Inequality 29 was taken from Heckert and Filliben (2003) and uses the student-t distribution.

$$|S_i - S_j| > t_{1-\alpha/2*,(b-1)(k-1)} \sqrt{\frac{2b\,(A-B)}{(b-1)\,(k-1)}} \tag{29}$$

Inequality 30 was taken from the manual of STATService 2.0 (Parejo *et al.*, 2012) and uses the standard normal distribution.

$$\left| \frac{W_i}{n_i\,(n_i+1)\,/2} - \frac{W_j}{n_j\,(n_j+1)\,/2} \right| > z_{1-\alpha/2*} \sqrt{\frac{k\,(k+1)\,(2n+1)\,(k-1)}{18n\,(n+1)}} \tag{30}$$

with

$$W_j = \sum_{i=1}^{b} (Q_i * R_{i,j}) \tag{31}$$

The calculated p-values can be adjusted to control Type I error inflation with the methods as given in `p.adjust`.

4.8 Example using `posthoc.quade.test`

```
> ## Conover (1999, p. 375f):
> ## Numbers of five brands of a new hand lotion sold in seven stores
> ## during one week.
> y <- matrix(c( 5,  4,  7, 10, 12,
+                1,  3,  1,  0,  2,
+               16, 12, 22, 22, 35,
+                5,  4,  3,  5,  4,
+               10,  9,  7, 13, 10,
+               19, 18, 28, 37, 58,
+               10,  7,  6,  8,  7),
+             nrow = 7, byrow = TRUE,
+             dimnames =
```

```
+             list(Store = as.character(1:7),
+                  Brand = LETTERS[1:5]))
> y

     Brand
Store  A  B  C  D  E
    1  5  4  7 10 12
    2  1  3  1  0  2
    3 16 12 22 22 35
    4  5  4  3  5  4
    5 10  9  7 13 10
    6 19 18 28 37 58
    7 10  7  6  8  7

> quade.test(y)

        Quade test

data:  y
Quade F = 3.8293, num df = 4, denom df = 24, p-value = 0.01519

> posthoc.quade.test(y, dist="TDist", p.adj="none")

        Pairwise comparisons using posthoc-Quade test with TDist approximation

data:  y

  A      B      C      D
B 0.2087 -      -      -
C 0.8401 0.2874 -      -
D 0.1477 0.0102 0.1021 -
E 0.0416 0.0021 0.0269 0.5172

P value adjustment method: none
```

5 Comparison of multiple joint samples (Two-factorial balanced incomplete block design)

5.1 Durbin test

The Durbin test can be applied in cases of an balanced incomplete block design (BIBD). The BIBD is characterized by

- Every block contains k groups (experimental units)

- Every group appears in r blocks
- Every group appears with every other group an equal number of times

The $H0$ is equivalent to the Friedman test or Quade test. First, rank the data $X_{i,j}$ within each block to obtain $R_{i,j}$. Then

$$R_j = \sum_{i=1}^{b} R_{i,j} \tag{32}$$

The test statistic is

$$\hat{\chi}^2 = \frac{(t-1)\left(\sum_{j=1}^{t} R_j^2 - r\ C\right)}{A - C} \tag{33}$$

with t the number of groups, k the total number of groups per block, b the number of blocks and r the number of times each group appears.

Furthermore,

$$\begin{aligned} A &= \sum_{i=1}^{b}\sum_{j=1}^{t} R_{i,j}^2 \\ C &= \frac{bk * (k+1)^2}{4} \end{aligned}$$

The $H0$ is rejected if $\hat{\chi}^2 > \chi^2$ on the level of α with $df = t - 1$.

It should be noted that the Durbin test is equivalent to the Friedman test in case of a balanced complete block design.

5.2 Example using `durbin.test`

```
> ## Example for an incomplete block design:
> ## Data from Conover (1999, p. 391).
> y <- matrix(c(
+ 2,NA,NA,NA,3, NA,  3,  3,  3, NA, NA, NA,  3, NA, NA,
+   1,  2, NA, NA, NA,  1,  1, NA,  1,  1,
+ NA, NA, NA, NA,  2, NA,  2,  1, NA, NA, NA, NA,
+   3, NA,  2,  1, NA, NA, NA, NA,  3, NA,  2,  2
+ ), ncol=7, nrow=7, byrow=FALSE,
+ dimnames=list(1:7, LETTERS[1:7]))
> y

   A  B  C  D  E  F  G
1  2  3 NA  1 NA NA NA
2 NA  3  1 NA  2 NA NA
3 NA NA  2  1 NA  3 NA
4 NA NA NA  1  2 NA  3
```

```
5  3 NA NA NA  1  2 NA
6 NA  3 NA NA NA  1  2
7  3 NA  1 NA NA NA  2

> durbin.test(y)

        Durbin rank sum test

data:  y
Durbin chi-squared = 12, df = 6, p-value = 0.06197
```

5.3 Durbin – posthoc test

Provided that the omnibus test of Durbin indicates, that at least one group differs from another group, multiple comparisons can be conducted according to Inequality 34

$$\|R_j - R_i\| > t_{1-\alpha/2*,bk-b-t+1} \sqrt{\frac{(A-C)\,2r}{bk-b-t+1}\left(1-\frac{\hat{\chi}^2}{b*(k-1)}\right)} \tag{34}$$

The p-values can be adjusted with `p.adjust`.

5.4 Example using `posthoc.durbin.test`

```
> posthoc.durbin.test(y, p.adj="none")

        Pairwise comparisons using Durbin's test for a two-way
                 balanced incomplete block design

data:  y

  A      B      C      D      E      F
B 0.4379 -      -      -      -      -
C 0.0114 0.0035 -      -      -      -
D 0.0035 0.0012 0.4379 -      -      -
E 0.0400 0.0114 0.4379 0.1411 -      -
F 0.1411 0.0400 0.1411 0.0400 0.4379 -
G 0.4379 0.1411 0.0400 0.0114 0.1411 0.4379

P value adjustment method: none
```

6 Auxiliary functions

The package `PMCMR` comes with a `print.PMCMR` and a `summary.PMCMR` function:

```
> print(posthoc.durbin.test(y, p.adj="none"))
```

```
        Pairwise comparisons using Durbin's test for a two-way
                 balanced incomplete block design

data:  y

  A      B      C      D      E      F
B 0.4379 -      -      -      -      -
C 0.0114 0.0035 -      -      -      -
D 0.0035 0.0012 0.4379 -      -      -
E 0.0400 0.0114 0.4379 0.1411 -      -
F 0.1411 0.0400 0.1411 0.0400 0.4379 -
G 0.4379 0.1411 0.0400 0.0114 0.1411 0.4379

P value adjustment method: none
```

```
> summary(posthoc.durbin.test(y, p.adj="none"))

        Pairwise comparisons using Durbin's test for a two-way
                 balanced incomplete block design

data:  y


P value adjustment method: none
        H0 statistic p.value
1  A  =  B 0.8164966  0.4379
2  A  =  C 3.2659863  0.0114
3  A  =  D 4.0824829  0.0035
4  A  =  E 2.4494897  0.0400
5  A  =  F 1.6329932  0.1411
6  A  =  G 0.8164966  0.4379
7  B  =  C 4.0824829  0.0035
8  B  =  D 4.8989795  0.0012
9  B  =  E 3.2659863  0.0114
10 B  =  F 2.4494897  0.0400
11 B  =  G 1.6329932  0.1411
12 C  =  D 0.8164966  0.4379
13 C  =  E 0.8164966  0.4379
14 C  =  F 1.6329932  0.1411
15 C  =  G 2.4494897  0.0400
16 D  =  E 1.6329932  0.1411
17 D  =  F 2.4494897  0.0400
18 D  =  G 3.2659863  0.0114
19 E  =  F 0.8164966  0.4379
```

```
20 E  =  G 1.6329932  0.1411
21 F  =  G 0.8164966  0.4379
```

Furthermore, the function `get.pvalues` was included, to extract the p-values from an `PMCMR` object or an `pairwise.htest` object. The output of `get.pvalues` is a named numeric vector, where each element is named after the corresponding pairwise comparison. It can be further processed with `multcompLetters` from the package `multcompView` to find and indicate homogeneous groups for a given level of significance. This can be used to create plots (Fig. 3) or tables (Table 1) with `xtable`.

```
> require(xtable)
> require(multcompView)
> data(InsectSprays)
> attach(InsectSprays)
> out <- posthoc.kruskal.dunn.test(count ~ spray, p.adjust="bonf")
> out.p <- get.pvalues(out)
> out.mcV <- multcompLetters(out.p, threshold=0.05)
> Rij <- rank(count)
> Rj.mean <- tapply(Rij, spray, mean)
> dat <- data.frame(Group = names(Rj.mean),
+                   meanRj = Rj.mean,
+                   M = out.mcV$Letters)
> dat.x <- xtable(dat)
> caption(dat.x) <- "Mean ranks ($\\bar{R}_{j}$) of the
+ \\texttt{InsectSprays} data set. Different letters (M)
+ indicate significant differences ($p < 0.05$) according
+ to the Bonferroni-Dunn test (see Chap. \\ref{Dunn})."
> colnames(dat.x) <- c("Group", "$\\bar{R}_{j}$", "M")
> digits(dat.x) <- 1
> label(dat.x) <- "tab:bonf.dunn"
> print(dat.x, include.rownames=F, caption.placement="top",
+                    sanitize.text.function = function(x){x},
+               table.placement="h")


> detach(InsectSprays)
```

Table 1: Mean ranks ($\bar{R}_j$) of the `InsectSprays` data set. Different letters (M) indicate significant differences ($p < 0.05$) according to the Bonferroni-Dunn test (see Chap. 2.4).

Group	$\bar{R}_j$	M
A	52.2	a
B	54.8	a
C	11.5	b
D	25.6	b
E	19.3	b
F	55.6	a

References

Conover WJ (1999). *Practical Nonparametric Statistics*. 3rd edition. Wiley.

Conover WJ, Iman RL (1979). "On multiple-comparison procedures." *Technical report*, Los Alamos Scientific Laboratory.

Demsar J (2006). "Statistical comparisons of classifiers over multiple data sets." *Journal of Machine Learning Research*, **7**, 1–30.

Dunn OJ (1964). "Multiple comparisons using rank sums." *Technometrics*, **6**, 241–252.

Glantz SA (2012). *Primer of biostatistics*. 7 edition. McGraw Hill, New York.

Heckert NA, Filliben JJ (2003). "Dataplot Reference Manual, Volume 2: Let Subcommands and Library Functions." *Technical Report 148*, National Institute of Standards and Technology.

Jonckheere AR (1954). "A distribution-free k-sample test against ordered alternatives." *Biometrica*, **41**, 133–145.

Nemenyi P (1963). *Distribution-free Multiple Comparisons*. Ph.D. thesis, Princeton University.

Parejo JA, García J, Ruiz-Cortés A, Riquelme JC (2012). "STATService: Herramienta de análisis estadístico como soporte para la investigación con Metaheurísticas. Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados."

R Core Team (2013). *R: A language and environement for statistical computing*. Vienna, Austria. URL `http://www.R-project.org/`.

Sachs L (1997). *Angewandte Statistik*. 8 edition. Springer, Berlin.

Siegel S, Castellan Jr NJ (1988). *Nonparametric Statistics for The Behavioral Sciences*. 2nd edition. McGraw-Hill, New York.

Wilcoxon F, Wilcox RA (1964). *Some rapid approximate statistical procedures.* Lederle Laboratories, Pearl River.

```
> require(multcompView)
> data(InsectSprays)
> attach(InsectSprays)
> out <- posthoc.kruskal.dunn.test(count ~ spray, p.adjust="bonf")
> out.p <- get.pvalues(out)
> out.mcV <- multcompLetters(out.p, threshold=0.05)
> Rij <- rank(count)
> Rj.mean <- tapply(Rij, spray, mean)
> xx <- barplot(Rj.mean, ylim=c(0, 1.2* max(Rj.mean)),
+               xlab="Spray", ylab="Mean rank")
> yy <- Rj.mean + 3
> text(xx, yy, lab=out.mcV$Letters)
> detach(InsectSprays)
```

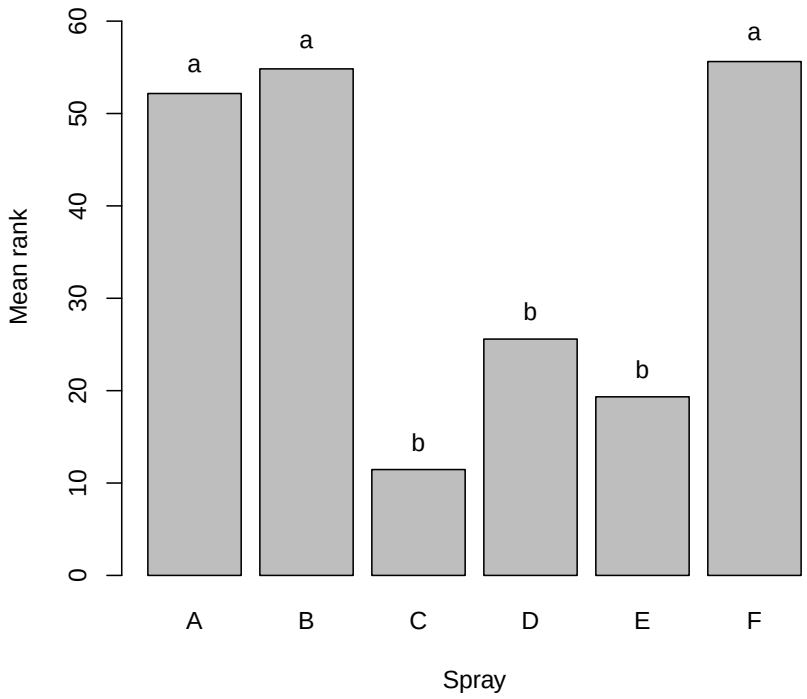

Figure 3: Barplot of mean ranks of the `InsectSprays` data set. Different letters indicate significant differences ($p < 0.05$) according to the Bonferroni-Dunn test (see Chap. 2.4).