

The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)

Thorsten Pohlert

June 24, 2014

© Thorsten Pohlert. This work is licensed under a Creative Commons License (CC BY-ND 4.0). See <http://creativecommons.org/licenses/by-nd/4.0/> for details. Please cite this package as: T. Pohlert (2014). *The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR)*. R package. See also `citation("PMCMR")`.

Contents

| | | |
|----------|--|----------|
| 1 | Introduction | 1 |
| 2 | Comparison of multiple independent samples (One-factorial design) | 2 |
| 2.1 | Kruskal and Wallis test | 2 |
| 2.2 | Kruskal-Wallis – post-hoc tests after Nemenyi | 2 |
| 2.3 | Examples using <code>posthoc.kruskal.nemenyi.test()</code> | 3 |
| 3 | Comparison of multiple joint samples (Two-factorial unreplicated complete block design) | 6 |
| 3.1 | Friedman test | 6 |
| 3.2 | Friedman – post-hoc test after Nemenyi | 6 |
| 3.3 | Example using <code>posthoc.friedman.nemenyi.test()</code> | 7 |

1 Introduction

For one-factorial designs with samples that do not meet the assumptions for one-way-ANOVA (i.e., i) errors are normally distributed, ii) equal variances among the groups, and, iii) uncorrelated errors) and subsequent post-hoc tests, the Kruskal-Wallis test (`kruskal.test`) can be employed that is also referred to as the Kruskal–Wallis one-way analysis of variance by ranks. Provided that significant differences were detected by the Kruskal-Wallis-Test, one may be interested in applying post-hoc tests for pairwise multiple comparisons of the ranked data. 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 (`friedman.test`). The consequent post-hoc pairwise multiple comparison test according to Nemenyi is also provided 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:

$$y_i = \mu + \alpha_i + \epsilon_i, \quad (1)$$

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:

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

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

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

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

$$\hat{H}^* = \hat{H}/C \quad (4)$$

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 $\hat{H} > \chi_{k-1; \alpha}^2$. 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.

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

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 ($|\bar{R}_i - \bar{R}_j|$) 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, as given by inequality 6:

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

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 not implemented in the package `PMCMR`, but can e.g. be employed with a Bonferroni-type adjustment of α .

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

The function `kruskal.test` is provided by the package `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:

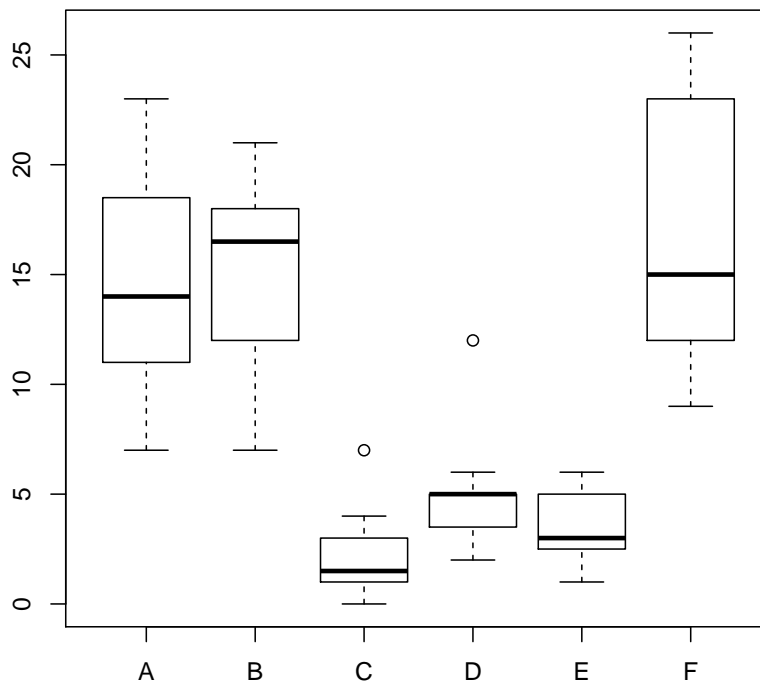
Based on a visual inspection, one can assume that the insecticides A, B, F differ from C, D, E . The global test can be conducted in this way:

```
> 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
```



Figure~1: Boxplot of the InsectSprays data set.

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()`. As the function has no `formula` enclosed, the response vector and the group vector have to be passed separately to the function.

```
> require(PMCMR)
> data(InsectSprays)
> attach(InsectSprays)
> posthoc.kruskal.nemenyi.test(x=count, g=spray, method="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 - - - -
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 $|\bar{R}_A - \bar{R}_C| : p < 0.01$. 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, method="Chisquare"))
```

```

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

```

data: count and spray

```

  A      B      C      D      E
B 0.9998 -      -      -      -
C 0.0004 0.0001 -      -      -
D 0.0860 0.0395 0.7427 -      -
E 0.0117 0.0041 0.9740 0.9909 -
F 0.9995 1.0000 6.8e-05 0.0307 0.0030

```

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

```

The test results can be aligned into a summary table as it is common in scientific articles. However, there is not yet a function included in the package PMCMR. Therefore, Table 1 was manually created.

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

3.1 Friedman test

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

$$y_{i,j} = \mu + \alpha_i + \pi_j + \epsilon_{i,j} \quad (7)$$

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, $H_0 : F(1) = F(2) = \dots = F(k)$ 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 $\pi_j : j = 1, \dots, m$. After that, the statistics of the Friedman test is calculated according to Eq. 8:

$$\hat{\chi}_R^2 = \left[\frac{12}{nk(k+1)} \sum_{i=1}^k R_i \right] - 3n(k+1) \quad (8)$$

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

3.2 Friedman – post-hoc test after Nemenyi

Provided that the Friedman test indicates significance, the post-hoc test according to Nemenyi can be employed (Sachs, 1997, p. 668). This test requires equal sample sizes ($n_1 = n_2 = \dots = n_k = n$) for each group k and a Friedman-type ranking of the data. The inequality 9 was taken from Demsar (2006, p. 11), where the critical difference refer to mean rank sums ($|\bar{R}_i - \bar{R}_j|$):

Table 1: Mean rank sums of insect counts (\bar{R}_i) after the application of insecticides (Group). Different letters indicate significant differences ($p < 0.05$) according to the Tukey-Kramer-Nemenyi post-hoc test. The global test according to Kruskal and Wallis indicated significance ($\chi^2(5) = 54.69, p < 0.01$).

| Group | \bar{R}_i | |
|-------|-------------|---|
| C | 11.46 | a |
| E | 19.33 | a |
| D | 25.58 | a |
| A | 52.17 | b |
| B | 54.83 | b |
| F | 55.62 | b |

$$|\bar{R}_i - \bar{R}_j| > \frac{q_{\infty;k;\alpha}}{\sqrt{2}} \sqrt{\frac{k(k+1)}{6n}} \quad (9)$$

This inequality (9) 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.

3.3 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 |

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

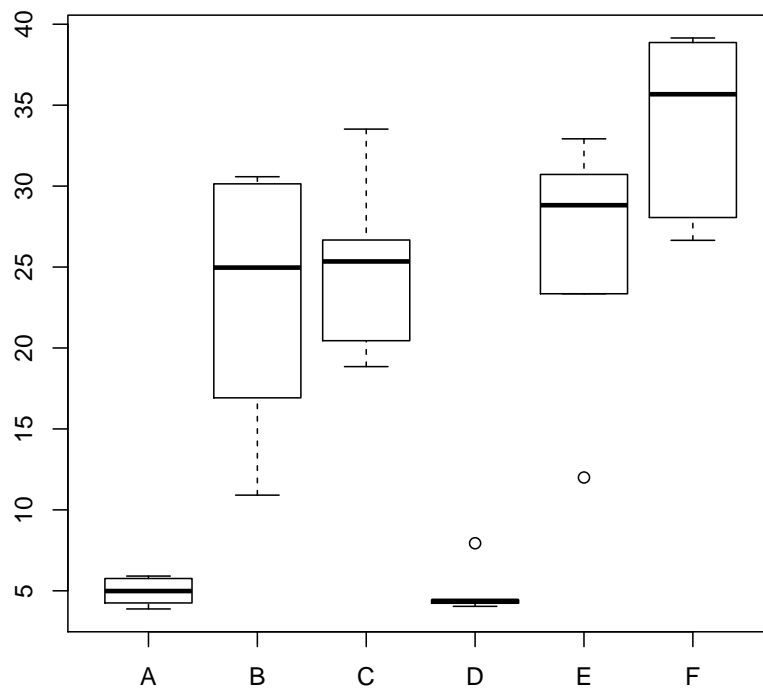
The global test is the Friedman test, that is already implemented in the package `stats` (R Core Team, 2013):

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

```

Friedman rank sum test

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



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

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 post-hoc test with q approximation for unrepli

```
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).

References

- Demsar J (2006). “Statistical comparisons of classifiers over multiple data sets.” *Journal of Machine Learning Research*, **7**, 1–30.
- R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria. URL <http://www.R-project.org/>.
- Sachs L (1997). *Angewandte Statistik*. 8 edition. Springer, Berlin.
- Wilcoxon F, Wilcox RA (1964). *Some rapid approximate statistical procedures*. Lederle Laboratories, Pearl River.