Research Article

Mario Hasler*

Multiple Contrasts for Repeated Measures

Abstract: This article addresses the problem of multiple contrast tests for repeated measures. Three procedures are described and compared by simulations concerning the familywise error type I. The procedure based on sandwich estimators seems to be most robust, except for all-pair comparisons.

Keywords: repeated measures, multiple contrasts, sandwich estimator, multivariate t-distribution, familywise error type I

***Corresponding author: Mario Hasler,** Christian-Albrechts-University, Kiel, Germany, E-mail: hasler@email.uni-kiel.de

1 Introduction

Multiple contrast tests (MCTs) and related simultaneous confidence intervals (SCIs) are well-known methods for testing and estimating linear functions of means called contrasts. A broad class of testing problems can be handled in specifying suitable contrast coefficients. The many-to-one comparison of Dunnett [1] is one of the most frequently applied and cited testing procedures today and it represents a simple example. Several treatments are compared with one control and tested for deviation. The all-pair comparison of Tukey [2], comparing all treatments against each other, is another very well-known example. Bretz [3] has formulated the trend test of Williams [4] as an approximate MCT. Here, the contrast coefficients depend in addition on the sample sizes of the treatment groups. Moreover, other problem-specific contrasts can be defined (see Nelson [5], Westfall [6] or Bretz et al. [7] for example). Furthermore, MCTs and SCIs can also be formulated for ratios of means (see Dilba et al. [8]) if conclusions about ratios – rather than differences – of means are of interest. This applies when relative changes are to be analyzed. Because correlations between the contrasts are involved by a joint distribution, MCTs exactly maintain the familywise error type I (FWE) over all contrasts. No further multiplicity adjustment is needed. SCIs (and adjusted p-values) are obtained for all hypotheses to be tested, complying with the guideline of the ICH [9]. Stepwise or gatekeeping procedures, for example, do not allow informative SCIs; see Strassburger and Bretz [10] and Guilbaud [11].

MCTs and related SCIs are usually confined to normally distributed, homoscedastic and independent data with one primary endpoint. For the heteroscedastic case, Hasler and Hothorn [12] described a solution based on Games and Howell [13]. Konietschke et al. [14] presented a non-parametric version. Hasler and Hothorn [15, 16] gave extensions for the case of multiple correlated endpoints. Unlike conventional MCTs, these approaches represent approximate solutions. They focus on situations where the usual assumptions are not fulfilled. Repeated measures also represent a situation where such an assumption is not met. For example, the means of several groups have to be compared simultaneously but the measurement values for these groups come from the same measurement objects (patients, plants, etc.), respectively. The groups are mostly according to time points but they can also represent different parts of the body, for example. If no type of mathematical function can be specified for the influence of the time points on the measurement values, a regression analysis is hardly appropriate or makes no sense. MCTs are intended instead. However, the measurement values are correlated and hence not independent. Ignoring these correlations would lead to over- or underestimation of treatment effects. As an example, see Section 5 where the aim is to identify the influence of the age of young boys in the context of a dental study.

This article presents three candidate procedures for handling the problem of repeated measures in the context of MCTs. In Section 2, the testing problem and first stochastic conclusions are described. The three procedures are introduced in Section 3, and they are compared by simulations studies in Section 4. An example is given in Section 5; Section 6 gives summary, discussion and outlook.

2 Testing problem

For $i = 1, \ldots, k$, $j = 1, \ldots, n$, let X_{ij} denote the jth observation at the ith time point in a one-way layout and $n - 1 \geq k$. The time points are regarded as factor levels. The vectors $\boldsymbol{X}_{\cdot j} = (X_{1j}, \ldots, X_{kj})'$ are mutually independent and follow a k-variate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots \mu_k)'$ and unknown covariance matrix $\Sigma = (\sigma_{ii'})_{i,i'} \in \mathbb{R}^{k \times k}$. This is,

$$\{X_{ij} : i = 1, \ldots, k\} \sim \perp N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (j = 1, \ldots, n).$$

Let $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_k)'$ and $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ii'})_{i,i'}$ be unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. The corresponding correlation matrix of the time points is denoted by $\boldsymbol{R} = (\rho_{ii'})_{i,i'} \in \mathbb{R}^{k \times k}$ with unbiased estimator $\hat{\boldsymbol{R}} = (\hat{\rho}_{ii'})_{i,i'}$.

The question of the trial may be, for example: at which time does the treatment cause the highest effect? Or is there a monotone trend over the entire time period? For that purpose, define the vector of contrasts $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)'$, where

$$\eta_l = \sum_{i=1}^{k} c_{li}\mu_i = c_l{}'\boldsymbol{\mu} \quad (l = 1, \ldots, q)$$

and $\hat{\eta}_l = \sum_{i=1}^{k} c_{li}\bar{X}_i = c_l{}'\bar{\boldsymbol{X}}$. The contrasts represent linear combinations of the means μ_i, where the vector c_l consists of contrast coefficients $c_{li} \in [-1, 1]$. A formulation based on ratios – instead of differences – of means following Dilba et al. [17] might also be possible. The aim is to simultaneously test the hypotheses

$$H_{0l} : \eta_l \leq \delta_l \quad (l = 1, \ldots, q), \tag{1}$$

where $\delta_l \in (-\infty, \infty)$ are contrast-specific thresholds. In many applications, $\delta_l = 0$ for all l. Testing problem [1] is a union–intersection test because the overall null hypothesis of interest, H_0, can be expressed as an intersection of the local null hypotheses, i.e.,

$$H_0 = \bigcap_{l=1}^{q} H_{0l}.$$

This means that H_0 is rejected if and only if at least one local null hypothesis H_{0l} is rejected.

Since

$$\mathrm{Var}(\hat{\eta}_l - \delta_l) = \frac{1}{n}\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{li}c_{li'}\sigma_{ii'} = \frac{1}{n}\, c_l{}'\boldsymbol{\Sigma}\mathbf{c}_l \quad (l = 1, \ldots, q),$$

it follows that, under H_{0l},

$$\frac{\sum_{i=1}^{k} c_{li}\bar{X}_i - \delta_l}{\frac{1}{n}\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{li}c_{li'}\sigma_{ii'}} = \frac{c_l{}'\bar{\boldsymbol{X}} - \delta_l}{\sqrt{\frac{1}{n}\, c_l{}'\hat{\boldsymbol{\Sigma}}\mathbf{c}_l}} \sim N(0, 1) \quad (l = 1, \ldots, q).$$

Consequently, the related test statistic for the lth contrast is

$$T_l = \frac{\sum_{i=1}^{k} c_{li}\bar{X}_i - \delta_l}{\sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{li}c_{li'}\hat{\sigma}_{ii'}}} = \frac{c_l{}'X - \delta_l}{\sqrt{\frac{1}{n}\, c_l{}'\hat{\Sigma}c_l}} \quad (l = 1, \ldots, q).$$

The marginal distribution of each T_l is a t-distribution with degree of freedom $df = n - 1$. The contrast coefficients c_{li} are usually applied to the means μ_i and their estimators $\bar{X}_i$. Because of the dependency of the observations X_{ij} due to the time points, however, the contrast coefficients c_{li} can even be applied to the measurement objects. This is, $\boldsymbol{c}_l{}'\boldsymbol{X}_{.1}, \ldots, \boldsymbol{c}_l{}'\boldsymbol{X}_{.n}$ can be regarded as a pseudo sample. Therefore, each single test statistic T_l represents a one-sample t-test for this pseudo sample. For example, in the simplest case of only $k = 2$ time points, a one-sample t-test of the differences $X_{21} - X_{11}, \ldots, X_{2n} - X_{1n}$ is usually applied.

In general, a joint distribution for $T_1, \ldots, T_q$ cannot be derived analytically so far. Only in the special case of homoscedastic and equicorrelated time points (compound symmetry), $T_1, \ldots, T_q$ follow a joint multivariate t-distribution. In general, however, the joint distribution is not among standard distributions discussed in textbooks. It is not a multivariate t-distribution, since the denominators of the different T_l represent different correlated χ^2 variables. A generalized t-distribution has been derived by Siddiqui [18] only in the bivariate case. An approximate multivariate extension in the equicorrelated case is given by Kotz et al. [19]; this approximation is exact in the bivariate case.

Nevertheless, under H_0 the test statistics $T_1, \ldots, T_q$ have a correlation matrix $\tilde{\boldsymbol{R}} = (\tilde{\rho}_{ll'})_{l,l'} \in \mathbb{R}^{q\times q}$ with elements

$$\begin{aligned}\tilde{\rho}_{ll'} &= \frac{\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{li}c_{l'i'}\sigma_{ii'}}{\sqrt{\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{li}c_{li'}\sigma_{ii'}}\sqrt{\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{l'i}c_{l'i'}\sigma_{ii'}}} \\ &= \frac{\boldsymbol{c}_l{}'\boldsymbol{\Sigma}\boldsymbol{c}_{l'}}{\sqrt{\boldsymbol{c}_l{}'\boldsymbol{\Sigma}\boldsymbol{c}_l}\sqrt{\boldsymbol{c}_{l'}{}'\boldsymbol{\Sigma}\boldsymbol{c}_{l'}}} \quad (l, l' = 1, \ldots, q).\end{aligned} \qquad [2]$$

This follows from

$$\mathrm{Cov}\left(\sum_{i=1}^{k} c_{li}\bar{X}_i, \sum_{i=1}^{k} c_{l'i}\bar{X}_i\right) = \frac{1}{n}\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{li}c_{l'i'}\sigma_{ii'} = \frac{1}{n}\,\boldsymbol{c}_{l'}{}'\boldsymbol{\Sigma}\boldsymbol{c}_{l'}.$$

As the covariances $\sigma_{ii'}$ are unknown, the estimators $\hat{\sigma}_{ii'}$ must be inserted into expression [2] instead. For the special case of independent time points, the covariance matrices $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$, respectively, have non-zero elements only on the diagonal. The test statistics and correlations then coincide with those of MCTs for heteroscedastic (independent) data (see for example [12]) in the balanced case.

3 Test procedures

The idea to use a multivariate t-distribution is one that should not be abandoned. It can be used as an approximation to the joint distribution of $T_1, \ldots, T_q$. The corresponding correlation matrix is the matrix $\tilde{\boldsymbol{R}}$ with elements given in eq. [2]. A naive approach uses the degree of freedom of the marginal t-distribution, namely

$$df = n - 1. \qquad [3]$$

This procedure is referred to here as the naive procedure. Although the covariances of the time points are taken into account, no further adjustment is achieved for the approximation of the joint distribution. As can be seen in Section 4, such a naive approach leads to liberal test decisions for small sample sizes. A smaller degree of freedom seems to be advisable. Such a candidate, investigated in the following, is

$$\tilde{df} = n - 1 - \frac{q-1}{k-1}. \qquad [4]$$

The related procedure is referred to here as the *df*-procedure.[1] It is exact for $k = 2$, as well as the naive procedure. Decreasing the degree of freedom means that higher (adjusted) p-values and absolute values for quantiles will be obtained.

A similar effect is reached if there is no adjustment of the (approximate) distribution of the test statistics, but if an adjustment is made of the related values of the test statistics themselves. This can be achieved by the use of sandwich estimators. The first mathematics in this context go back to Huber [20]. MacKinnon and White [21] introduced the HC3 estimator in the context of linear regression models in the presence of heteroscedasticity to improve the performance in small samples. A critical review of sandwich estimation, as it is applied today in many statistical areas, is given by Freedman [22]; see also Zeileis [23, 24] for an overview and for a realization in *R* [25]. Herberich et al. [26] applied the HC3 sandwich estimator to MCTs, especially related to Tukey [2], in the presence of heteroscedasticity. The HC3 sandwich estimator of σ_{ii} is given by

$$\hat{\sigma}_{ii}^{*} = \frac{1}{n}\sum_{j=1}^{n} \omega_{ij} \quad (i = 1, \ldots, k)$$

where

$$\omega_{ij} = \frac{\left(X_{ij} - \bar{X}_i\right)^2}{\left(1 - \frac{1}{n}\right)^2} \quad (i = 1, \ldots, k; j = 1, \ldots, n)$$

(see Zeileis [23]). It follows that

$$\begin{aligned}\hat{\sigma}_{ii}^{*} &= \frac{n}{(n-1)^2}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2 \\ &= \frac{n}{n-1}\,\hat{\sigma}_{ii} \quad (i = 1, \ldots, k).\end{aligned}$$

As the sample sizes are balanced for the time points, the sandwich estimation leads to a constant increase of the sample variances. Hence, the related sandwich estimator for the covariance matrix Σ is given by

$$\hat{\boldsymbol{\Sigma}}^{*} = \begin{pmatrix} \sqrt{\hat{\sigma}_{11}^{*}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\hat{\sigma}_{kk}^{*}} \end{pmatrix} \tilde{\boldsymbol{R}} \begin{pmatrix} \sqrt{\hat{\sigma}_{11}^{*}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\hat{\sigma}_{kk}^{*}} \end{pmatrix} = \frac{n}{n-1}\,\hat{\Sigma}.$$

1 This degree of freedom resulted from earlier simulation studies of the author on this topic. It turned out to have the best properties compared to other potential degrees of freedom.

The sandwich estimation leads to more robust (i.e. higher) absolute values for the sample (covariances). This causes smaller absolute values for the corresponding test statistics, given by

$$T_l^* = T_l\sqrt{\frac{n-1}{n}} = \frac{\sum_{i=1}^{k} c_{li}\bar{X}_i - \delta_l}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{k}\sum_{i'=1}^{k} c_{li}c_{li'}\hat{\sigma}_{ii'}}} = \frac{\boldsymbol{c}_l'\bar{\boldsymbol{X}} - \delta_l}{\sqrt{\frac{1}{n-1}\boldsymbol{c}_l'\hat{\boldsymbol{\Sigma}}\boldsymbol{c}_l}} \quad (l = 1, \ldots, q).$$

The correlation matrix of the test statistics $T_1^*, \ldots, T_q^*$ under H_0 is equal to $\tilde{\boldsymbol{R}}$ with elements as in eq. [2]. Again, as the covariances $\sigma_{ii'}$ in expression [2] are unknown, the related estimators must be inserted instead. However, inserting $\hat{\sigma}_{ii'}^*$ leads to the same result as inserting $\hat{\sigma}_{ii'}$, as the sample sizes are balanced for the time points.

Thus, use test statistics T_l^* instead of T_l, and use degree of freedom [3]. This approach suffers from the fact that the sample variances are biased; they are just asymptotically unbiased [21]. Consequently, it is not exact for the simplest case $k = 2$; conservative test decisions can be expected here. This procedure is referred to here as the sandwich procedure.

4 Simulation study

The procedures described in Section 3 are based on approximations by multivariate *t*-distributions. Simulations concerning the FWE were used to assess the quality of the procedures and to identify the best one in this regard. Several types of contrasts were considered which are all related to hypotheses [1]: Dunnett (many-to-one), Tukey (all-pair), Williams (trend), and Average (mean averages). For reasons of consistency, the Tukey contrasts were also tested one-sided. The number of time points, k, varied from 3 to 6. The (balanced) sample sizes per group were $n = (k+1) + \{0, 2, 4, 8, 16, 32\}$. The groups had means $\mu_i = 100$ and standard deviations $\sqrt{\sigma_{ii}} = 10$ (homoscedastic time points) or $\sqrt{\sigma_{ii}} = 10 + 30(i-1)/(k-1)$ (heteroscedastic time points), respectively ($i = 1, \ldots, k$). Three equicorrelation structures (compound symmetry) of the time points were chosen ($\rho^{\min}$, 0, 0.8) as well as a random correlation structure (different for each simulation run); $\rho^{\min} = -1/(k-1)$ denotes the minimal equicorrelation. The FWE was simulated at a nominal level of $\alpha = 0.05$. The simulation results were obtained from 10,000 simulation runs each, with starting seed 123456, using a program code in the statistical software R [25], applying package mvtnorm [27, 28]. A (linear) *p*-value interpolation had to be implemented for the *df*-procedure. This was because the function pmvt of the package mvtnorm [27, 28] calculates *p*-values of multivariate *t*-distributions just for integer degrees of freedom. Real values for degrees of freedom are usually rounded down. This would cause very conservative test decisions for the *df*-procedure when the sample size is very small.

The results of the simulations are compressed and reduced to avoid an inflation of the length of this article. The complete results can be obtained from the author on request. Hence, Figures 1–4 show the FWE of the three procedures only for $k = 3$ and $k = 6$ time points, respectively. The different correlation structures and the two variance situations are not explicitly shown because they had no notable influence on the results. They are summarized so that they constitute the boxplots. Each figure represents one of the four types of contrasts. The respective three upper (lower) sub-figures are always related to $k = 3$ ($k = 6$) time points; the respective two left (middle, right) sub-figures are always related to the naive (*df*-, sandwich) procedure. The results of the simulations can be summarized as follows. The quality of the procedures differs depending on type of contrast, number of time points and sample size. In general, the naive

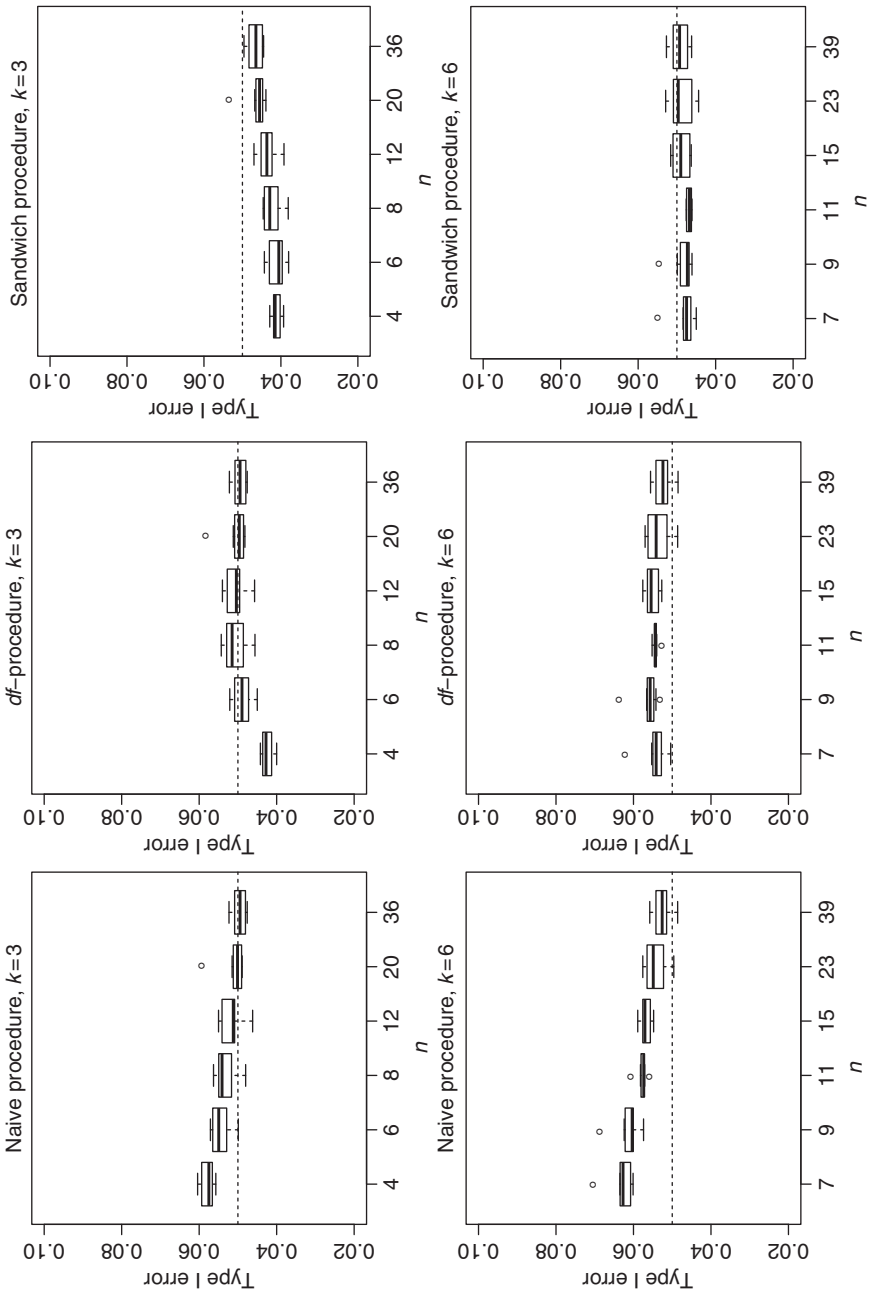

Figure 1 Global type I error for $k = 3$ and $k = 6$ time points, respectively, Dunnett contrasts, several correlations and variances.

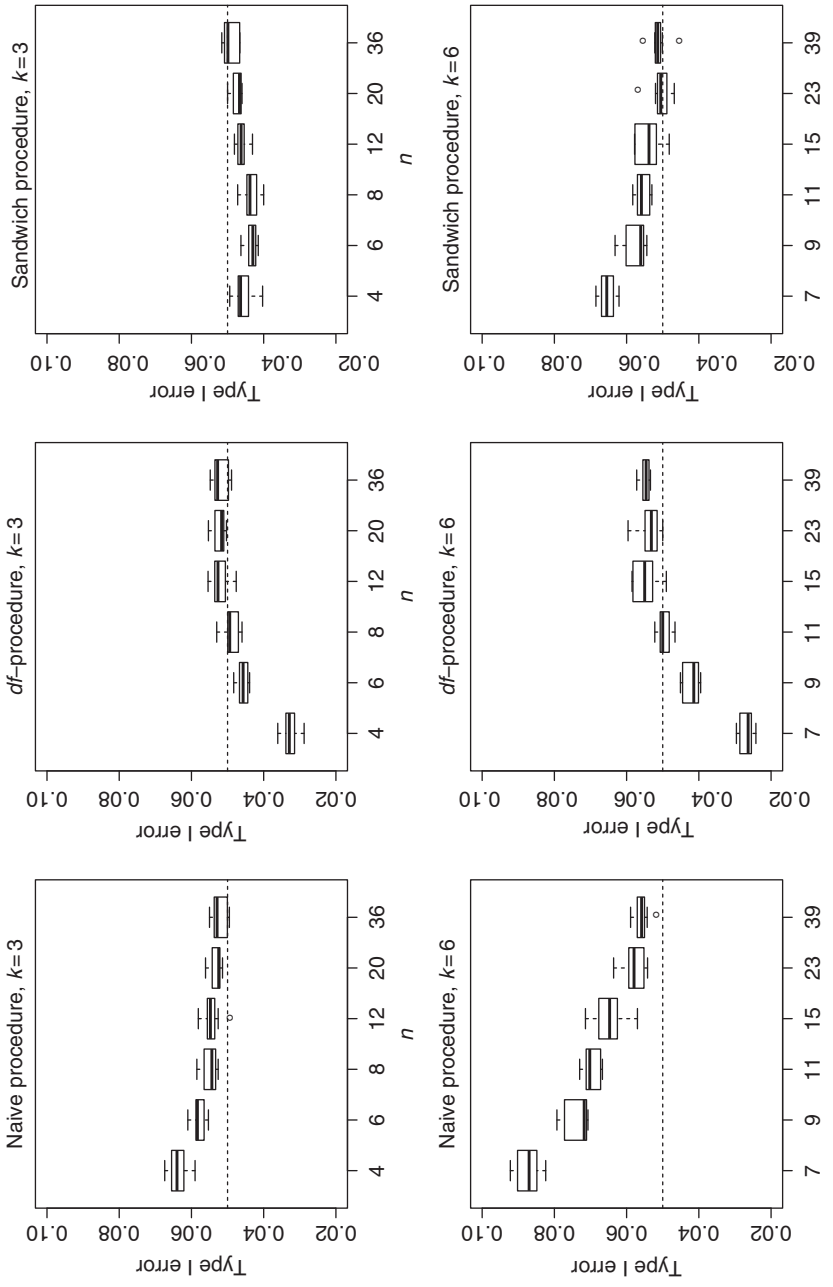

Figure 2 Global type I error for $k = 3$ and $k = 6$ time points, respectively, Tukey contrasts, several correlations and variances.

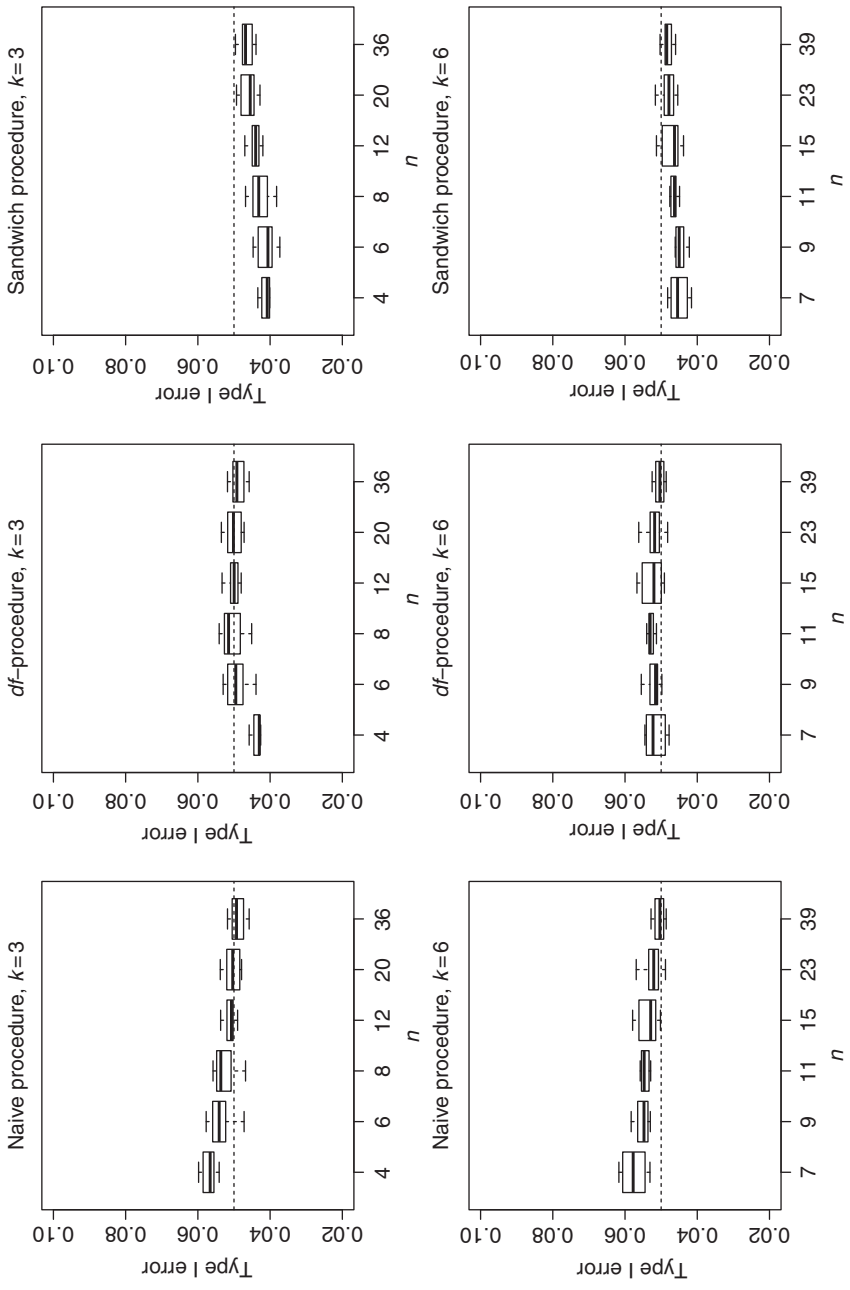

Figure 3 Global type I error for $k = 3$ and $k = 6$ time points, respectively, Williams contrasts, several correlations and variances.

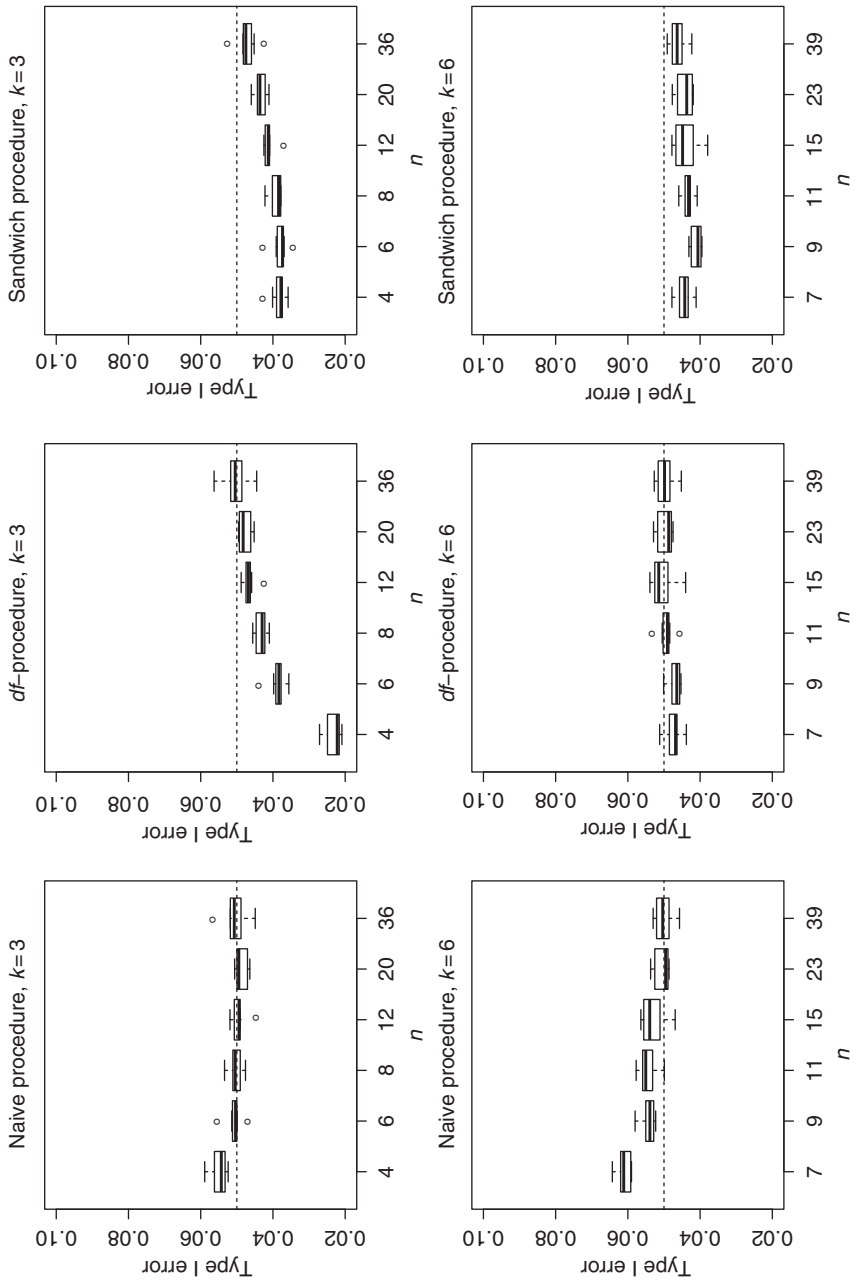

Figure 4 Global type I error for $k = 3$ and $k = 6$ time points, respectively, Average contrasts, several correlations and variances.

procedure is liberal for almost all sample sizes considered. It becomes more liberal for decreasing sample size, increasing number of time points and especially for Tukey contrasts. Because of the smaller degree of freedom, the *df*-procedure is much less liberal than the naive one and even conservative for small sample size and small number of time points, depending on the type of contrast. It becomes less conservative, up to liberal, for increasing number of time points, except for Tukey contrasts, where it is generally conservative for small sample sizes. The sandwich procedure is conservative in general. This conservatism vanishes for increasing number of time points and sample size, except for Tukey contrasts, where it becomes liberal for increasing number of time points.

The influence of the type of contrast is not surprising since different types of contrasts cause different numbers of comparisons and different correlations among the contrasts. Tukey contrasts seem to be an exceptional case, however. The number of all-pair comparisons exhibits quadratic growth because $q = k(k-1)/2 = \binom{k}{2}$, whereas $q = k-1$ for Dunnett or Williams contrast, and $q = k$ for Average contrasts. The ratio of the number of contrasts to the number of time points seems to have influence. This fact was attempted to be taken into account by the *df*-procedure; it is the reason for degree of freedom [4].

The influence of the sample size is clear. The higher the sample size, the less the three procedures differ because their FWE converges to the nominal level α. This is obvious for the naive and the *df*-procedure but also for the sandwich procedure, since the sandwich estimator is asymptotically unbiased. In general, an increase of the number of time points reduces potential conservatism, except for the *df*-procedure with Tukey contrasts. Moreover, if a procedure exceeds the nominal level α, this effect is intensified by an increasing number of time points.

5 Example

Potthoff and Roy [29] published data of a dental study at the University of North Carolina Dental School on 11 girls and 16 boys at ages 8, 10, 12 and 14. Each measurement is the distance (in millimetres) between the centre of the pituitary and the pterygomaxillary fissure. The complete data set can be loaded from package nlme [30] of the statistical software R [25]. However, only the measurements of the boys are considered here. This data subset can also be loaded from R-package nparLD [31]. Both the statistical software R [25] and the packages are available at http://www.r-project.org. Figure 5 shows the related boxplots per age. An

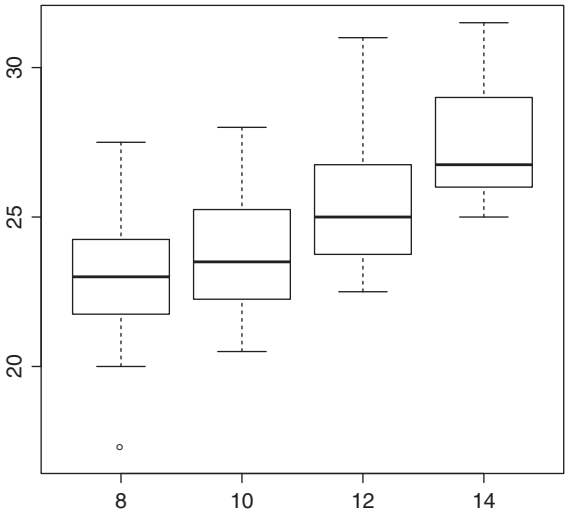

Figure 5 Boxplot of the dental data set of Potthoff and Roy [29].

experimental question for these data could be: do the distances increase over the time? Then a many-to-one comparison related to Dunnett [1] but for repeated measures is appropriate with age 8 as the control group.

The differences of interest are

$$\eta_l = \mu_{l+1} - \mu_1 \quad (l = 1, 2, 3),$$

and the hypotheses to be tested are given by

$$H_{0l} : \eta_l \leq 0 \quad (l = 1, 2, 3).$$

The estimated covariance matrix, used by the naive and the *df*-procedure, is

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 6.017 & 2.292 & 3.629 & 1.613 \\ 2.292 & 4.563 & 2.194 & 2.810 \\ 3.629 & 2.194 & 7.032 & 3.241 \\ 1.613 & 2.810 & 3.241 & 4.349 \end{pmatrix},$$

whereas the sandwich procedure uses the estimation

$$\hat{\mathbf{\Sigma}}^* = \begin{pmatrix} 6.418 & 2.444 & 3.871 & 1.720 \\ 2.444 & 4.867 & 2.340 & 2.998 \\ 3.871 & 2.340 & 7.501 & 3.457 \\ 1.720 & 2.998 & 3.457 & 4.639 \end{pmatrix}.$$

As expected, all elements of $\hat{\mathbf{\Sigma}}^*$ are greater than those of $\hat{\mathbf{\Sigma}}$. They differ by factor $16/15 = 1.067$.

Table 1 shows the results of the data evaluation for the three procedures. There is a significant difference in means compared to age 8 for ages 12 and 14, but not for age 10. The adjusted *p*-values do not differ very much, but the naive procedure has the smallest, reflecting the liberal behaviour of that procedure. As expected, the sandwich procedure has smaller values of the test statistics compared to the naive and the *df*-procedure, whereas the *df*-procedure has a smaller degree of freedom compared to naive and the sandwich procedure.

Table 1 Summary of the test for the dental data set of Potthoff and Roy [29] according to the different procedures.

Comparison	Procedure	Statistic	Degree of freedom	(adj.) *p*-value
10–8	naive	1.5315	15	0.1510
	df-	1.5315	14.33	0.1521
	sandwich	1.4828	15	0.1626
12–8	naive	4.7270	15	0.0003
	df-	4.7270	14.33	0.0005
	sandwich	4.5769	15	0.0005
14–8	naive	6.8764	15	< 0.0001
	df-	6.8764	14.33	< 0.0001
	sandwich	6.6580	15	< 0.0001

6 Summary and discussion

Independence of measurements is a usual assumption of MCTs. It is fulfilled if the several groups of a trial represent treatments, for example, and if each group consists of different measurement objects. If the

groups represent different time points or parts of the body, however, all groups normally consist of the same measurement objects. Hence, observations are repeated and correlated. MCTs must be adapted then so that they also take these correlations into account. The joint distribution of related test statistics is not among the standard distributions discussed in textbooks. Three procedures have been considered in this article, representing approximate solutions based on multivariate *t*-distributions. The naive approach generally leads to liberal test decisions. The procedure applying an adjusted degree of freedom can tend to both liberal and conservative test decisions, whereas the procedure based on sandwich estimators maintains the FWE, except for all-pair comparisons. Simulations concerning the power have not been carried out because it is not useful to do a power comparison for procedures which are known not to maintain the FWE in some situations. The sandwich procedure can obviously be expected to have a high power for all-pair comparisons and a lower power else.

The assumptions made in Section 2 are restricted to observations from one treatment group measured at several time points (for example). The related question of the trial may be, for example: at which time does the treatment cause the highest effect? MCTs have been investigated in this article for repeated measures, i.e. correlated measures, in a one-way layout. However, in a typical practical situation, there is often not just one but many treatments involved in the study, measured at several time points. This is actually a two-way layout, where the measurements are independent with regard to the treatments and dependent with regard to the time points. Several practical questions and related statistical hypotheses then arise. The question "at which time does the treatment cause the highest effect" can be asked for each treatment simultaneously. The procedures considered in this article had to be extended to this context. On the other hand, one could be interested in a comparison of the treatments simultaneously for the time points. Comparison of treatment effects on a single time point follow the well-known theory of MCTs. Stochastic problems occur due to the simultaneous strategy for the multiple, correlated time points. A simple solution is the adoption of MCTs for multiple endpoints according to Hasler and Hothorn [15], where the time points are simply regarded as endpoints.

The procedures described in this article still have the assumptions that the data are normally distributed and come from a completely randomized design. If data are not normally distributed, if they come from more complex experimental designs (block design, split plot design,...), or if covariates are included, the specific procedures described cannot be used. However, the general approaches of adjusted degrees of freedom or sandwich estimators for covariance matrices can also be adopted in generalized linear models [32], mixed models [33, 34] or ANCOVA models [35] respectively. The related adjustments of the *df*- and the sandwich procedure are only based on the number of time points (or other dependent groups), number of contrasts and the sample size, respectively. An application to the latter models could be a next, interesting step. For example, it is well-known that current multiple comparisons based on mixed models also yield liberal or conservative test decisions depending on the degree of freedom approximation (see for example Faes et al. [36]). For a non-parametric test procedure in the context of repeated measures, see Konietschke et al. [37] for example.

References

1. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 1955;50:1096–121.
2. Tukey JW. The problem of multiple comparisons. Dittoed manuscript of 396 pages. Department of Statistics, Princeton University, Princeton, NJ, 1953.
3. Bretz F. An extension of the Williams trend test to general unbalanced linear models. Comput Stat Data Anal 2006;50:1735–48.
4. Williams DA. A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 1971;27:103–17.
5. Nelson PR. Multiple comparisons of means using simultaneous confidence intervals. J Qual Technol 1989;21:232–41.

6. Westfall PH. Multiple testing of general contrasts using logical constraints and correlations. J Am Stat Assoc 1997;92:299–306.
7. Bretz F, Hothorn T, Westfall P. Multiple comparisons using R. London: Chapman & Hall/CRC, 2011.
8. Dilba G, Bretz F, Guiard V, Hothorn LA. Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. Methods Inf Med 2004;43:465–9.
9. ICH E9 Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials. Stat Med 1999;18:1905–42.
10. Strassburger K, Bretz F. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. Stat Med 2008;27:4914–27.
11. Guilbaud O. Note on simultaneous inferences about non-inferiority and superiority for a primary and a secondary endpoint. Biom J 2011;53:927–37.
12. Hasler M, Hothorn LA. Multiple contrast tests in the presence of heteroscedasticity. Biom J 2008;50:793–800.
13. Games PA, Howell JF. Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. J Educ Stat 1976;1:113–25.
14. Konietschke F, Hothorn LA, Brunner E. Rank-based multiple test procedures and simultaneous confidence intervals. Electron J Stat 2012;6:738–59.
15. Hasler M, Hothorn LA. A Dunnett-type procedure for multiple endpoints. Int J Biostat 2011;7:3.
16. Hasler M, Hothorn LA. A multivariate Williams-type trend procedure. Stat Biopharm Res 2012;4:57–65.
17. Dilba G, Bretz F, Guiard V. Simultaneous confidence sets and confidence intervals for multiple ratios. J Stat Plan Inference 2006;136:2640–58.
18. Siddiqui MM. A bivariate t distribution. Ann Math Stat 1967;38:162–6.
19. Kotz S, Balakrishnan N, Johnson NL. Continuous multivariate distributions, 2nd ed. New York: John Wiley and Sons, 2000.
20. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, 1967:221–33.
21. MacKinnon JG, White H. Some heteroskedasticity-consistent covariance-matrix estimators with improved finite-sample properties. J Econometrics 1985;29:305–25.
22. Freedman DA. On the so-called "Huber sandwich estimator" and "Robust standard errors." Am Statistician 2006;60:299–302.
23. Zeileis A. Econometric computing with HC and HAC covariance matrix estimators. J Stat Softw November 2004;11(10):1–17.
24. Zeileis A. Object-oriented computation of sandwich estimators. J Stat Softw August 2006;16(9):1–16.
25. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012. Available at: http://www.R-project.org/. ISBN 3-900051-07-0.
26. Herberich E, Sikorski J, Hothorn T. A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. Plos One 2010;5:e9788.
27. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. mvtnorm: multivariate normal and t distributions, 2012. Available at: http://CRAN.R-project.org/package=mvtnorm, R package version 0.9–9994.
28. Hothorn T, Bretz F, Genz A. On multivariate t and Gauss probabilities in R. R News 2001;1:27–9.
29. Potthoff RF, Roy SN. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika 1964;51:313–26.
30. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: linear and nonlinear mixed effects models, 2012. Available at: http://CRAN.R-project.org/package=nparLD, r package version 3.1–104.
31. Noguchi K, Latif M, Thangavelu K, Konietschke F, Gel YR, Brunner E. nparLD: nonparametric analysis of longitudinal data in factorial experiments, 2012. Available at: http://CRAN.R-project.org/package=nparLD, r package version 2.0.
32. McCullagh P, Nelder JA. Generalized linear models, 2nd ed. London: Chapman & Hall/CRC, 1989.
33. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–74.
34. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer, 2000.
35. Cochran WG. Analysis of covariance – its nature and uses. Biometrics 1957;13:261–81.
36. Faes C, Molenberghs G, Aerts M, Verbeke G, Kenward MG. The effective sample size and an alternative small-sample degrees-of-freedom method. Am Stat 2009;63:389–99.
37. Konietschke F, Bathke AC, Hothorn LA, Brunner E. Testing and estimation of purely nonparametric effects in repeated measures designs. Comput Stat Data Anal 2010;54:1895–905.