Comparisons of multiple treatment groups with a negative control or placebo group: Dunnett test vs. closed test procedure

Ludwig A. Hothorn,
Im Grund 12, D-31867 Lauenau, Germany
(retired from Leibniz University Hannover)

December 9, 2020

Abstract

Several treatments are usually compared with a control using the Dunnett test. As an alternative, three variants of the closed testing approach are considered, one with ANOVA-F-tests, one with MCT-GrandMean and one with global Dunnett-tests in the partition hypotheses.

1 The problem

Comparisons of multiple treatment groups with a control (or placebo) group are frequently performed in biomedical experiments: in a plant molecular experiment new mutants were compared with the wild type [24], in a toxicological assay selected clinical chemistry endpoints in three dose groups with disodium adenosine-triphosphate were compared with a water control [13] and in neurological clinical trial three candidate drugs amiloride, fluoxetine, and riluzole were compared with a placebo group for the primary endpoint volumetric MRI percentage brain volume change [6].

Typically, the Dunnett procedure is used [7] in such randomized one-way designs. The main advantage is the availability of simultaneous confidence intervals, alternatively to the multiplicity-adjusted p-values. An alternative is the closed testing procedure (CTP) [15], because it is available for any test in the GLM (particularly its small $n_i$ approximations [19]) and can be easily generalized to further joint analysis, e.g. considering multiple endpoints [5]. Three test versions are considered here for the global and partition hypotheses: the common ANOVA F-test, the global multiple contrast test for comparisons against the grand mean [18] and the global Dunnett test. Since the last two are based on multiple contrast tests, subset contrasts can be formulated for all k-samples, allowing the entire degree of freedom to be used in all sub-tests.

2 The Dunnett procedure

In the following, a low-dimensional one-way layout $y_{ij} = \mu + \text{factor}_i + \epsilon_{ij} (i = 0, \ldots, k)$ with $\epsilon_{ij} \sim N(\mu, \sigma^2)$ is considered. The Dunnett test can be formulated as multiple contrast test (MCT) [20], [17]:

$t_{MCT} = \max(t_1, \ldots, t_q)$ with $t_q = \sum_{i=0}^k c_i \bar{y}_i / S \sqrt{\sum_{i=0}^k c_i^2 / n_i}$ where $c_i$ are the contrast coefficients (see below). The common-used adjusted p-values are given by the minimum empirical $\alpha$-level fulfilling...
the equality \( \frac{1}{\sqrt{1+\xi_i}} \frac{c_i}{n_i} = l_q \cdot df \cdot R \cdot 1 - sided, \alpha \), where \( l_q \cdot df \cdot R \cdot 1 - sided, \alpha \) is the quantile of central q-variate t distribution, available in the package mvtnorm [16]. Compatible to the adjusted p-values are (two) or one-sided simultaneous confidence limits which are not considered here because of their difficulties in the CTP [8][23].

3 The closed testing procedure for the comparison of treatment groups with a control

An important issue of the CTP [15] is the definition of the interesting elementary hypotheses, i.e. exactly those which are to be interpreted. Adjusted p-values are available for exactly these. In the above described design, the many-to-one hypotheses \( H_i: \mu_i - \mu_0 \) are of interest only. Starting from this, one defines all subset intersection hypotheses up to the global hypothesis, involving these hypotheses. One rejects \( H_i \) at level \( \alpha \) if and only if \( H_i \) itself is rejected and all hypotheses which include them (each at level \( \alpha \)). Each hypothesis is tested with a level \( \alpha \) test, with any appropriate test - this allows a high flexibility of the here described approach. Each of these tests (determined by the \( \xi \) elementary hypotheses) is an intersection-union test (IUT), i.e. \( T_{\xi} = \min(T_1, ..., T_\xi) \), or more common \( \rho_{CTP} = \max(p_1, ..., p_\xi) \). In general CTP, the subset hypotheses can be complex and contradictory, but when considering hypotheses for comparisons with a control, they form a simple, so-called complete family of hypotheses [21] (same as with the 2-sample multiple endpoint problem). Using the monotonicity of the p-values and the dependencies of the sub-hypotheses, shortcuts are possible in general CTP [4]. By using additive or p-value combination tests, a variety of procedures can be constructed [10] in the general case.

For the simple design with \( k = 2 \), the family include the elementary (e.g. \( H_0^{[1]} \)), intersection (e.g. \( H_0^{[12]} \)) and global hypotheses (e.g. \( H_0^{[123]} \)):

- \( H_0^{[1]}: \mu_0 = \mu_1 \subset [H_0^{[12]}, H_0^{[13]}] \subset H_0^{[123]} \)
- \( H_0^{[2]}: \mu_0 = \mu_2 \subset [H_0^{[12]}, H_0^{[23]}] \subset H_0^{[123]} \)
- \( H_0^{[3]}: \mu_0 = \mu_3 \subset [H_0^{[13]}, H_0^{[23]}] \subset H_0^{[123]} \)

3.1 Using ANOVA F-tests

In the global, partition and elementary hypotheses the common omnibus F-test can be used: \( F_{\text{omnibus}} = \frac{\sum_{i=1}^{k} \sum_{j=0}^{n_i} (y_{ij} - \bar{y}_j)^2 / n_i}{\sum_{i=0}^{k} \sum_{j=0}^{n_j} (y_{ij} - \bar{y})^2 / (N-k-1)} \sim F_{k,N-k-1} \). These global (partial) homogeneity null hypotheses \( H^i: \mu_i = ... = \mu_k \) are compared with \( H_{\text{F-test}}: \mu_i \neq \mu'_i \), for at least one i, any i, i \( \neq \) i' for \( (k-1)/2 \) pairwise comparisons. This means that for the global hypothesis an F-test is used for all \( (k + 1) \) groups, for the first stage of the partition hypothesis an F-test is used for \( (k + 1) / 2 \) groups only, and so on, up to the elementary hypothesis an F-test with only 2 groups (0, i). In the elementary hypotheses this loss of df can be avoided by using a pairwise contrast test for \( (k + 1) \) groups (the so called multiple t-tests)-remember, one is free in the choice of the tests in CTP.

3.2 Using global MCT for comparisons against grand mean

The multiple contrast test against the grand mean (MCT-GM) represents an alternative to the F-test as global omnibus test [14] although the alternative is different from those of the F-test: \( H^i_{\text{MCT-GM}}: \mu_i \neq \mu_0 \), i.e. only \( k \) comparisons are considered. (Notice, this global test version ignores a key advantage of MCT-GM here: the availability of simultaneous confidence intervals.) The MCT-GM is a special case of a multiple contrast test, a maximum test: \( t_{\text{MCT}} = \max(l_1, ..., l_q) \) where the q contrast tests...
Contrast $t_{\text{Contrast}} = \sum_{i=0}^{k} c_i \bar{y}_i / \sqrt{\sum_{i=0}^{k} c_i^2 / n_i}$ and the specific contrast coefficients (here again for $k=3+1$) in the design with a control (C) and three treatments ($T_{1,2,3}$):

| $c_i$ | C | $T_1$ | $T_2$ | $T_3$ |
|------|----|------|------|------|
| $c_a$ | -1 | 1/3  | 1/3  | 1/3  |
| $c_b$ | 1/3 | -1   | 1/3  | 1/3  |
| $c_c$ | 1/3 | 1/3  | -1   | 1/3  |
| $c_d$ | 1/3 | 1/3  | 1/3  | -1   |

Table 1: Contrast matrix for global grand mean comparisons

The here considered global MCT-GM is defined by $p_{\text{MCT-GM}}^{\text{global}} : \max(p_{ii'})$. The advantage of this modified MCT-GM is the use of the full df for all hypotheses but subset contrasts for partition and elementary hypotheses. The contrast matrix for $H_{012}^{012}$ is for example:

| $c_i$ | C | $T_3$ | $T_2$ | $T_3$ |
|------|----|------|------|------|
| $c_a$ | -1 | 1/2  | 1/2  | 0    |
| $c_b$ | 1/2 | -1   | 1/2  | 0    |
| $c_c$ | 1/2 | 1/2  | -1   | 0    |

Table 2: Subset contrast matrix for grand mean comparisons for $H_{012}^{012}$

The power advantage of MCT-GM with respect to the F-test can be observed for small $n_i$ and several patterns of the alternative [18], where particularly the least favorable configuration (LFC) $[0, 0, \ldots, 0, \xi(k+1)/k]$ (effect size $\xi$) is of interest [14]. A major advantage of the MCT’s over the F-test is the easy availability of one-sided tests to avoid directional errors [23].

### 3.3 Using global Dunnett tests

In the global and partition hypotheses global Dunnett test can be used: $p_{\text{Dunnett}}^{\text{global}} = \max(p_1, \ldots, p_q)$. Accordingly two-sample t-tests are used in the elementary hypotheses. This power difference is a complex association between the $n_i$ (especially for unbalanced designs), the shape of the alternative, the effect size and the dimension $k$. Notice, in the elementary hypotheses again pairwise contrast tests for $(k+1)$ groups are used. Again, subset contrast matrices are used to keep the entire df of the $k$-sample design.

| $c_i$ | C | $T_3$ | $T_2$ | $T_3$ |
|------|----|------|------|------|
| $c_a$ | -1 | 0    | 0    | 1    |
| $c_b$ | -1 | 0    | 1    | 0    |
| $c_c$ | -1 | 1    | 0    | 0    |

Table 3: Global Dunnett-type contrast matrix

Notice, the adjusted p-values for the Dunnett original procedure, the global Dunnett-test and the Grand Mean MCT were estimated by means of the package multcomp [12].

### 4 Simulation study

In a simulation study for random experiments with a single primary endpoint $y_{ij}$, $k = 3 + 1$, considering balanced and $n_0 = \text{sqrt}(n_i)$ unbalanced one-way design and normal homoscedastic errors $\epsilon_{ij} \sim N(\mu, \sigma^2)$ selected alternatives, among them $[\mu_0 = \mu_1 = \mu_2 < \mu_3]; [\mu_0 = \mu_1 < \mu_2 < \mu_3]; [\mu_0 <$
\( \mu_1 = \mu_2 < \mu_3; [\mu_0 < \mu_1 < \mu_2 = \mu_3], \) were compared according to their FWER and any-pairs power (5000/2000 runs). Common simulation studies on MCTs compare the any-pair power \([11]\) or average power \([22]\) only. These concepts simplifies power comparisons considerably, but are not target-oriented, since which comparison is in the alternative is not considered. But one does not want to know if any mutants differ from the wild type, or any dose from the negative control, or any therapy vs. placebo. No, you want to evaluate exactly a particular mutant, dose or exactly a selected therapy relative to control (see the motivating examples above). Therefore the concept of per-pairs power is shown in the Appendix, although it is difficult to interpret (and therefore \(k=3+1\) was used). The four tests are abbreviated with \(D\) (Dunnett original), CTP-Du (subset Dunnett global), CTP-Fi (CTP ANOVA), and CTP-GM (subset CTP GrandMean) with \(D_1\) the pairwise power, \(D^\alpha\) the any-pairs power. Instead complete power curves, only a relevant points in the alternative with \(\pi \approx 0.8\) is considered

\[
\begin{array}{cccccc}
 n_i & \mu_j & \text{Dunnett} & \text{CTP-Du} & \text{CTP-Fi} & \text{CTP-GM} \\
\hline
(8,5,5,5) & 0, 0, 0, \delta & 0.87 & 0.91 & 0.86 & 0.88 \\
& 0, 0, \delta, \delta & 0.76 & 0.76 & 0.78 & 0.83 \\
& 0, \delta, \delta, \delta & 0.76 & 0.78 & 0.70 & 0.67 \\
& 0, \delta, \delta, \delta & 0.76 & 0.87 & 0.86 & 0.81 \\
& 0, \delta, \delta, \delta & 0.77 & 0.77 & 0.71 & 0.74 \\
& 0, \delta, \delta, \delta & 0.77 & 0.70 & 0.70 & 0.70 \\
& 0, 0, \delta, \delta & 0.80 & 0.80 & 0.77 & 0.81 \\
& 0, \delta, \delta, \delta & 0.79 & 0.83 & 0.77 & 0.75 \\
& 0, \delta, \delta, \delta & 0.76 & 0.87 & 0.86 & 0.81 \\
& 0, \delta, \delta, \delta & 0.78 & 0.79 & 0.72 & 0.75 \\
& 0, \delta, \delta, \delta & 0.79 & 0.80 & 0.73 & 0.73 \\
& 0, 0, \delta, \delta & 0.91 & 0.92 & 0.91 & 0.92 \\
& 0, 0, \delta, \delta & 0.83 & 0.83 & 0.87 & 0.88 \\
& 0, \delta, \delta, \delta & 0.81 & 0.83 & 0.78 & 0.74 \\
& 0, \delta, \delta, \delta & 0.81 & 0.84 & 0.92 & 0.88 \\
& 0, \delta, \delta, \delta & 0.81 & 0.81 & 0.78 & 0.78 \\
& 0, \delta, \delta, \delta & 0.81 & 0.82 & 0.78 & 0.76 \\
& 0, \delta, \delta, \delta & 0.94 & 0.97 & 0.97 & 0.97 \\
& 0, \delta, \delta, \delta & 0.84 & 0.84 & 0.83 & 0.86 \\
& 0, \delta, \delta, \delta & 0.84 & 0.87 & 0.83 & 0.80 \\
& 0, \delta, \delta, \delta & 0.83 & 0.94 & 0.95 & 0.90 \\
& 0, \delta, \delta, \delta & 0.85 & 0.85 & 0.79 & 0.81 \\
& 0, \delta, \delta, \delta & 0.84 & 0.85 & 0.81 & 0.79 \\
& 0, \delta, \delta, \delta & 0.92 & 0.95 & 0.92 & 0.93 \\
& 0, 0, \delta, \delta & 0.83 & 0.84 & 0.86 & 0.88 \\
& 0, \delta, \delta, \delta & 0.82 & 0.84 & 0.80 & 0.75 \\
& 0, \delta, \delta, \delta & 0.83 & 0.93 & 0.93 & 0.90 \\
& 0, \delta, \delta, \delta & 0.80 & 0.80 & 0.79 & 0.78 \\
& 0, \delta, \delta, \delta & 0.84 & 0.84 & 0.81 & 0.78 \\
& 0, \delta, \delta, \delta & 0.95 & 0.95 & 0.98 & 0.98 \\
& 0, 0, \delta, \delta & 0.82 & 0.83 & 0.81 & 0.84 \\
& 0, \delta, \delta, \delta & 0.82 & 0.87 & 0.84 & 0.83 \\
& 0, \delta, \delta, \delta & 0.83 & 0.93 & 0.93 & 0.90 \\
& 0, \delta, \delta, \delta & 0.82 & 0.83 & 0.77 & 0.78 \\
& 0, \delta, \delta, \delta & 0.82 & 0.84 & 0.80 & 0.79 \\
\end{array}
\]

Table 4: Any pairs power of the 4 tests (bold ... higher, italic... lower power in CTP-GM vs. Dunnett)

Per definition all tests controls FWER empirically in both weak and strong sense (see the Appendix). Five tendencies reveal: i) no umpt exists, ii) the CTP-Du reveals high power for \([\mu_0 < \mu_1 = \ldots = \mu_{k-1} = \mu_k]\) alternative, iii) CTP-GM reveals high power for the \([\mu_0 = \mu_1 = \ldots = \mu_{k-1} < \mu_k]\) alternative, i.e. the least favorable configuration of the Dunnett Test \([9]\), iv) lowest power for CTP-GM is for alternatives where several groups contributes in proportion to the non-centrality, and v) for extreme unbalanced designs (e.g. \(n_i = 8, 5, 5, 2\)) and plateau shape the power loss of the Dunnett tests is clearly exceeded by the CTP tests.
5 Evaluation of a data example

A randomized dose finding trial with 4 doses of a new compound (and placebo) to treat the irritable bowel syndrome measuring the primary efficacy endpoint baseline adjusted abdominal pain score was selected [3]. (The raw data are available as data(IBScovars) in library(DoseFinding) [2].)

![Boxplot for abdominal pain example](image)

In the original work, the intention was a stratified analysis for males and females; here, gender is modeled as an additive factor without interaction. Although larger values of the endpoint represent a clinical benefit, two-sided tests are used for reasons of comparability. The two-sided multiplicity adjusted p-values for the 4 elementary hypotheses $\mu_i - \mu_0$ are given in Table 5.

| Comparison | Dunnett original | CTP-F | CTP-Du | CTP-GM |
|------------|------------------|-------|--------|--------|
| 5-1        | 0.0234           | 0.0346| 0.0222 | 0.0121 |
| 4-1        | 0.0229           | 0.0346| 0.0222 | 0.0117 |
| 3-1        | 0.0654           | 0.0346| 0.0358 | 0.0226 |
| 2-1        | 0.0779           | 0.0346| 0.0358 | 0.0234 |

Table 5: Adjusted p-values for abdominal pain score

The sensitivity advantages of all CTP’s are obvious in this example. This even leads to qualitatively different statements: all doses are significantly different compared to placebo. The CTP-GM is particularly sensitive. The decision trees can be displayed particularly well with the CRAN package CPT [1] in Figure 2.

![CTP's: (left) F-test, (middle) Dunnett global test, (right) GrandMeans MCT](image)
6 Summary

Three modifications of the closed test procedure for comparisons of multiple treatment groups with a control as an alternative to the Dunnett procedure are proposed. As marginal tests the ANOVA F-Test, the global Dunnett test and the global test for Grand Mean multiple contrasts are considered. The disadvantage of unavailable confidence intervals is contrasted by the power advantage for selected shapes of the alternative, where of course none umpt exists. Related R-software is available on example style. Extensions in the generalized linear (mixed) model, ratio-to-control inference, variance heterogeneity and non-parametric testing will be reported soon.

References

[1] CTP: Closed Testing Procedure. Authors Paul Jordan and Herpers Matthias. 2020. R package version 2.0.0, url = https://CRAN.R-project.org/package=CTP.
[2] DoseFinding: Planning and Analyzing Dose Finding Experiments, Author Bjoern Bornkamp, 2020. R package version 0.9-17, url = https://CRAN.R-project.org/package=DoseFinding.
[3] E. Biesheuvel and L. A. Hothorn. Many-to-one comparisons in stratified designs. Biometrical Journal, 44(1):101–116, 2002.
[4] W. Brannath and F. Bretz. Shortcuts for locally consonant closed test procedures. Journal of the American Statistical Association, 105(490):660–669, June 2010.
[5] F. Bretz, W. Maurer, W. Brannath, and M. Posch. A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine, 28(4):586–604, February 2009.
[6] J. Chatawy, F. De Angelis, P. Connick, R. A. Parker, D. Plautone, A. Doshi, N. John, J. Stutters, D. MacManus, F. P. Carrasco, F. Burkof, S. Ourselin, M. Brasher, M. Ross, G. Cranswick, S. H. Pavitt, G. Giovannoni, C. A. G. Wheeler-Kingshott, C. Hawkins, B. Sharrack, R. Bastow, C. J. Weir, N. Stallard, S. Chandran, T. Williams, T. Beyene, V. Bassan, A. Zapata, P. Connick, D. Lyle, J. Cameron, D. Mollison, S. Colville, B. Dhillon, A. Walker, L. Smith, S. Gnanapavan, R. Nicholas, W. Rashid, J. Aram, H. Ford, J. Overell, C. Young, H. Arndt, J. Guadagno, N. Evangelou, M. Craner, J. Palace, J. Hobart, B. Sharrack, D. Paling, C. Hawkins, S. Kalra, and B. McLean. Efficacy of three neuroprotective drugs in secondary progressive multiple sclerosis (ms-smart): a phase 2b, multiarm, double-blind, randomised placebo-controlled trial. Lancet Neurology, 19(3):214–225, March 2020.
[7] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50(272):1096–1121, 1955.
[8] O. J. M. Guilbaud. Simultaneous confidence intervals compatible with sequentially rejective graphical procedures. Statistics in Biopharmaceutical Research, 10(3):220–232, 2018.
[9] A. J. Hayter and W. Liu. A method of power assessment for tests comparing several treatments with a control. Communications in Statistics- Theory and Methods, 1992.
[10] K. S. S. Henning and P. H. Westfall. Closed testing in pharmaceutical research: Historical and recent developments. Statistics in Biopharmaceutical Research, 7(2):126–147, April 2015.
[11] L. A. Hothorn and R. Pirow. Use compatibility intervals in regulatory toxicology. Regulatory Toxicology and Pharmacology, 116:104720, October 2020.
[12] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. Technical report, Biometrical J, 2008.
[13] R. Jager, M. Purpura, and J. C. Fuller. Subchronic (90-day) repeated dose toxicity study of disodium adenosine-5’-triphosphate in rats. Regulatory Toxicology and Pharmacology, 116:104760, October 2020.
[14] F. Konietzschke, S. Bosiger, E. Brunner, and L. A. Hothorn. Are multiple contrast tests superior to the anova? International Journal of Biostatistics, 9(1):63–73, May 2013.
[15] R. Marcus, E. Peritz, and K. R. Gabriel. Closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63(3):655–660, 1976.
[16] X. F. Mi, T. Miwa, and T. Hothorn. mvtnorm: New numerical algorithm for multivariate normal probabilities. R Journal, 1(1):37–39, May 2009.
[17] H. Makerjee, T. Robertson, and F. T. Wright. Comparison of several treatments with a control using multiple contrasts. Journal of the American Statistical Association, 82(399):902–910, September 1987.
[18] P. Pallmann and L. A. Hothorn. Analysis of means: a generalized approach using r. Journal of Applied Statistics, 43(8):1541–1560, June 2016.
7 Appendix
Table 6: FWER and per-pairs power. D (Dunnett original), N (Dunnett global), C (CTP ANOVA), W (CTP MCT-GrandMean)