Closed test procedures for the comparison of dose groups against a negative control group or placebo

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

January 1, 2021

Abstract

Dose groups are compared with a control assuming an order restriction usually by the Williams trend test. Here, as an alternative, two variants of the closed testing procedure are considered, one where global Williams tests are used in the partition hypotheses, and another where pairwise contrast tests are used for this purpose. Related R software is provided.

1 The problem

Comparisons of $k$ dose groups with a negative control assuming a monotonic dose-response relationship are often performed in biomedical experiments by means of the Williams trend test [16]. For example, organ weights of rats in three concentrations of pentabromodiphenyl ethers in a toxicological bioassay [3] or the reduction of serum uric acid in gout-free subjects considering two doses of tuna extract compared to placebo [8] in a randomized clinical trial. One uses the Williams test instead of the Dunnett test [2] (without the assumption of an order restriction) because of the higher power (due to the restriction of $H_1$) and especially the possibility to interpret a trend (both globally and locally for selected parts of the dose-response relationship). The main difference between the Dunnett test and the Williams test is that the former considers comparisons between $C$ and individual $D_i$, but the latter does not consider comparisons with explicit doses, only pooled doses (except $D_{\max} - C$). Therefore, an order-restricted test is derived here for the comparison to the control with the individual doses. The closed test procedure [9] (CTP) is an alternative to the max-T test for multiple contrasts on which the Williams test is based [1]. Two special cases of CTP are considered here: the complete hypothesis family when comparing to control alone [14] and the decision tree reduction when assuming an order restriction [5]. Thus, related closed test versions of order-restricted tests are derived here.

2 A brief description of the Williams procedure

In the original paper [16] the procedure was described for maximum likelihood estimators under order restriction, but one can formulate this simplified as multiple contrast test (MCT) [13]. The basis is a maximum test over several t-distributed standardized contrast tests $t_q$ (here referred to as maxT-test): $t_{\text{MCT}} = \max(t_1, \ldots, t_q')$ with $t_q = \sum_{i=0}^k c_i \bar{y}_i / S \sqrt{\sum_{i=0}^k c_i^2 / n}$, where $c_i$ are the contrast coefficients (see selected examples below). A adjusted $p$-values are given by the minimum empirical $\alpha$-level:
\[ \frac{\sum_{i=0}^{k} c_i y_i}{\sum c_i^2/n_i} = t_{q,df,R,1-sided,1-\alpha} \] where \( t_{q,df,R,1-sided,1-\alpha} \) is the quantile of central q-variate t-distribution, easily available in the package mvtnorm [10]. Compatible to the adjusted p-values should be simultaneous confidence intervals. They are not considered here because of their difficulties in general closed tests [4].

3 A brief description of restricted closed testing procedures

Starting point of any CTP is the a problem-adequate definition of the interesting elementary hypotheses, here: \( H_i : \mu_i - \mu_0 \). In a second step a decision tree containing all subset intersection hypotheses up to the global hypothesis, involving these elementary hypotheses is constructed [9]. \( H_i \) is rejected at level \( \alpha \) if and only if \( H_i \) itself is rejected and all hypotheses which include them in the decision tree (again 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) represents an intersection-union test (IUT), i.e. \( T^{\xi} \), or more common \( p^{CTP} = \max(p_1, ..., p_k) \). Here this approach is demonstrated for a rather simple design with \( k = 2 \), the family include the following elementary (e.g. \( H_0^{123} \)), intersection (e.g. \( H_0^{012} \)) and global hypotheses (e.g. \( H_0^{0123} \)):

- \( H_0^{01} : \mu_0 = \mu_1 \subset [H_0^{012}, H_0^{013}] \subset H_0^{0123} \)
- \( H_0^{02} : \mu_0 = \mu_2 \subset [H_0^{012}, H_0^{023}] \subset H_0^{0123} \)
- \( H_0^{03} : \mu_0 = \mu_3 \subset [H_0^{013}, H_0^{023}] \subset H_0^{0123} \)

Monotonic order constraint \( H_1 : \mu_0 \leq \mu_1 \leq ... \leq \mu_k | \mu_0 < \mu_k \) (for any possible pattern of equalities/inequalities) further greatly simplifies this specific CTP. Under this restriction of \( H_1 \), rejection of \( H_0^{0123} \) implies rejection of \( H_0^{013} \) and \( H_0^{03} \), and rejection of \( H_0^{012} \) implies rejection of \( H_0^{02} \), and so on. The hypothesis system is highly simplified:

- \( H_0^{01} : H_0^{01} \wedge H_0^{012} \wedge H_0^{0123} \)
- \( H_0^{02} : H_0^{012} \wedge H_0^{0123} \)
- \( H_0^{03} : H_0^{023} \)

For these hypotheses, any one-sided level \( \alpha \) test can be used. The elementary hypotheses should be tested with contrast tests for \( \mu_i - \mu_0 \), not with two-sample tests, so as not to reduce power in small \( n_i \) designs. For the partition and global hypothesis, any order-constrained test can be used. Because of the \( D_i - C \) comparison, two versions of the special CTP are considered here: i) Williams global test for each subset (denoted CW), ii) pairwise contrast tests for \( \mu_i - \mu_0 \) (where \( \xi \) is the highest dose in each subset) (denoted CP).

4 Simulation study

Empirical power and size of these tests are demonstrated by a tiny simulation study for a low-dimensional one-way design \( y_{ij} = \mu + \text{factor}_i + \epsilon_{ij} (i = 0, ..., k) \) with \( \epsilon_{ij} \sim N(\mu_i, \sigma^2) \). Random experiments with a single primary endpoint \( y_{ij}, k = 2 \) in a balanced design were used, estimating their per-pair power \( \pi_{01}, \pi_{02}, \pi_{03} \) for six strictly monotonic alternatives and two shapes with a downturn effect at the high dose (occurring in some in-vitro toxicity assays). Common simulation studies in the framework of simultaneous inference compare any-pair power [6] or average power [15]. These concepts greatly simplify the power comparisons, but are not purposeful because they do not take into account which individual comparison is currently in the alternative. However, one does not want to know whether any dose is different from
the negative control. No, one wants to evaluate exactly a particular dose relative to the control. That is why the concept of per-pair power is used here, although it is difficult to interpret (and that is why \(k = 2\) was used).

The four tests are abbreviated as \(D\) (Dunnett original), \(W3\) (Williams \(D_3 - C\) contrast only, since the other contrasts are non-pairwise), \(CW\) (CTP using a subset of Williams global tests), and \(CP\) (CTP using pairwise contrasts), where \(D_i\) is pairwise power (\(D^n\) is any-pairs power for the Dunnett test for reference purpose). Instead of complete power curves, only one relevant point in the alternative is considered for about \(\pi > 0.8\).

| Shape | \(H_0\) | \(H_1\) | \(D_1\) | \(D_2\) | \(D_3\) | \(W_3\) | CW1 | CW2 | CW3 | CP1 | CP2 | CP3 |
|-------|--------|--------|--------|--------|--------|--------|------|------|------|------|------|------|
| \(\mu_0 = \mu_1 = \mu_2 = \mu_3\) | \(\mu_1 < \mu_2\); \(\mu_3 \geq \mu_2\); \(\delta\) | 0.10 0.82 0.81 0.85 | 0.18 0.86 0.95 | 0.18 0.84 0.90 |
| Monot | \(\mu_0 = \mu_1 = \mu_2 < \mu_3\); \(\delta\) | 0.42 0.80 0.81 0.85 | 0.57 0.87 0.95 | 0.54 0.83 0.91 |
| \(\mu_0 < \mu_1\); \(\mu_2 < \mu_3\) | 0.02 0.02 0.81 | 0.02 0.05 0.85 | (not shown) |
| Non-m | \(\mu_0 < \mu_1\); \(\mu_2 < \mu_3\) | 0.09 0.42 0.80 0.85 | (not shown) |

Table 1: Per-comparison power estimates for selected alternatives

By definition, all tests control the familywise error rate (not shown in detail here). For strictly monotonic alternatives, the power of the Williams test is by definition slightly greater than that of the Dunnett test (directly comparable only for \(D_3 - 0\)). Both CTP-tests almost always show superiority in power for all \(\pi_i\), and for some patterns a marked superiority over Dunnett’s test. As expected, the power of the CW test is slightly better than that of the CP test, but the latter being impressive for its simplicity. Depending on the magnitude of the response decline in \(D_{max}\), all tests assuming an order restriction are not robust, as expected.

5 Evaluation of a data example

Relative liver weights in male rats of 4 dose groups and a negative control (abbreviated as 1) of an in vivo bioassay are used as a data example (data available in library(nparcomp)). The boxplots in Figure 1 show an approximately symmetrical distribution and homogeneous variances, so the standard tests are used.

Figure 1: Jittered boxplots for liver weights

The one-sided multiplicity adjusted p-values for the 4 elementary hypotheses \(\mu_1 - \mu_0\) are given in Table 2. As expected, the p-value \(p_{5-1}\) is smaller in the Williams test and both closure tests than in the Dunnett test. The p-values for the 4 elementary hypotheses \(\mu_1 - \mu_0\) are smaller in CP than in CW, not surprising for such a step-wise shape of the dose-response relationship. p-values for the elementary
Comparison | Dunnett | Williams | CW | CP
--- | --- | --- | --- | ---
5-1 | 1.99 $^{-0}$ | 8.8 $^{-1}$ | 8.8 $^{-1}$ | 4.7 $^{-1}$
4-1 | 1.51 $^{-3}$ | NA | 7.1 $^{-4}$ | 4.3 $^{-4}$
3-1 | 0.531 | NA | 0.322 | 0.253
2-1 | 0.719 | NA | 0.406 | 0.406

Table 2: Adjusted p-values for liver weight data (NA... not available)

hypotheses $[\mu_4 - \mu_1], [\mu_3 - \mu_1], [\mu_2 - \mu_1]$ are not available per definition for the Williams-type test. The R-code of this example is given in the Appendix.

6 Conclusions

Obviously, no uniformly powerful test can exist for any pattern of monotonic $H_1$, certainly not for alternatives with downturns at high dose(s). In particular, if one considers the specific pattern of dose-response as a priori unknown, the CTP’s proposed here prove to be a powerful alternative. In particular, the availability of adjusted p-values for the elementary hypotheses makes these tests attractive. Further generalizations for generalized linear mixed effect model (glm) (e.g., for proportions [7]), use for estimating the no-observed-adverse-event-level (NOAEL), or consideration of trend tests for modeling dose as a quantitative covariate [12], and a software implementation will follow shortly.

References

[1] F. Bretz. An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics and Data Analysis*, vol. 50, no. 7, Art. no. 7, 2006.
[2] 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.
[3] J. K. Dunnick, K. R. Shockley, A. R. Pandiri, G. E. Kissling, K. E. Gerrish, T. V. Ton, R. E. Wilson, S. S. Brar, A. E. Brix, S. Waidyanatha, E. Mutlu, and D. L. Morgan. Pbde-47 and pbde mixture (de-71) toxicities and liver transcriptomic changes at pnd 22 after in utero/postnatal exposure in the rat. *Archives of Toxicology*, 92(11):3415–3433, November 2018.
[4] O. J. M. Guilbaud. Simultaneous confidence intervals compatible with sequentially rejective graphical procedures. *Statistics in Biopharmaceutical Research*, 10(3):220–232, 2018.
[5] L. A. Hothorn, M. Neuhauser, and H. F. Koch. Analysis of randomized dose-finding-studies: Closure test modifications based on multiple contrast tests. *Biometrical Journal*, 39(4):467–479, 1997.
[6] L. A. Hothorn and R. Prow. Use compatibility intervals in regulatory toxicology. *Regulatory Toxicology and Pharmacology*, 116:104720, October 2020.
[7] L. A. Hothorn Comparisons of proportions in k dose groups against a negative control assuming order restriction: Williams-type test vs. closed test procedures. [arXiv:2011.13758] November 2020.
[8] D. Kubomura, M. Yamada, and A. Masui. Tuna extract reduces serum uric acid in gout-free subjects with insignificantly high serum uric acid: A randomized controlled trial. *Biomedical Reports*, 5(2):254–258, August 2016.
[9] 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.
[10] X. F. Mi, T. Miwa, and T. Hothorn. mvtnorm: New numerical algorithm for multivariate normal probabilities. *R Journal*, 1(1):37–39, May 2009.
[11] H. Mukerjee, 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.
[12] F. Schaarschmidt, C. Ritz, and L. A. Hothorn. The R library tukeytrend. 2020.
[13] J. P. Shaffer. Multiple comparisons emphasizing selected contrasts - extension and generalization of Dunnett’s procedure. *Biometrics*, 33(3):293–303, 1977.
[14] E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, 50(5):641–656, October 2008.
library(nparcomp)
library(multcomp)
data(liver)

liver$dose <- as.factor(liver$dosage)
mod1 <- lm(weight ~ dose, data = liver)
CM04 <- c(-1,0,0,0,1)
CM03 <- c(-1,0,0,1,0)
CM02 <- c(-1,1,0,0,0)
CM01 <- c(-1,1,1,0,0)
ni <- aggregate(weight ~ dose, data = liver, length)$total
Cmat0123 <- contrast(mat = ni[1:4], type = "Williams"); V4 <- c(0,0,0)
Cmat012 <- contrast(mat = ni[1:3], type = "Williams"); V3 <- c(0,0)
Cmat01 <- contrast(mat = ni[1:2], type = "Williams"); V2 <- c(0)

T04 <- summary(glht(mod1, linfct = mcp(dose = CM04), alternative = "greater"))$test$pvalues
T03 <- summary(glht(mod1, linfct = mcp(dose = CM03), alternative = "greater"))$test$pvalues
T02 <- summary(glht(mod1, linfct = mcp(dose = CM02), alternative = "greater"))$test$pvalues
T01 <- summary(glht(mod1, linfct = mcp(dose = CM01), alternative = "greater"))$test$pvalues
W01234 <- min(summary(glht(mod1, linfct = mcp(dose = "Williams"), alternative = "greater"))$test$pvalues)
W0123 <- min(summary(glht(mod1, linfct = mcp(dose = Cmat0123), alternative = "greater"))$test$pvalues)
W012 <- min(summary(glht(mod1, linfct = mcp(dose = Cmat012), alternative = "greater"))$test$pvalues)
W01 <- min(summary(glht(mod1, linfct = mcp(dose = Cmat01), alternative = "greater"))$test$pvalues)

CTP4 <- T04
CTP3 <- max(T04, T03)
CTP2 <- max(T04, T03, T02)
CTP1 <- max(T04, T03, T02, T01)

CTW4 <- W01234
CTW3 <- max(W01234, W0123)
CTW2 <- max(W01234, W0123, W012)
CTW1 <- max(W01234, W0123, W012, W01)

DT04 <- summary(glht(mod1, linfct = mcp(dose = "Dunnett"), alternative = "greater"))$test$pvalues
DTW4 <- summary(glht(mod1, linfct = mcp(dose = "Williams"), alternative = "greater"))$test$pvalues
