DR-IBS: AG Landwirtschaftliches Versuchswesen, Monheim Juni 29, 2023
Statistical design and analysis of field trials on non-target species
Alternatives to OECD-decision tree

Ludwig A. Hothorn
31867 Lauenau (retired from Leibniz University Hannover).
E-mail: *ludwig@hothorn.de*. Homepage: www.hothorn.de

June 22, 2023

Structure I

- Possible effect of new agro-chemicals on non-target species in randomized field trials $\Rightarrow$ small sample sizes issues inherently (not only power!)
- Design: i) CRD, ii) $n_i = 6$ pre-defined by OECD guidance (follow strictly!), iii) $C+, C_1, ..., C_3, C+$, iv) multiple sampling times T_0, T_t, v) many primary multiple count endpoints (mostly solicited, unsolicited) and their taxonomic aggregations (serious multiplicity issue!)
- Actually a proof of safety, i.e. non-inferiority tests (clearly a directional hypothesis) or better confidence limits. But species-specific tolerable thresholds ξ_p unknown. Therefore proof of hazard - with all the nice confusions, e.g. *'The absence of evidence is no evidence of absence'* (Altman/Bland)

An example I

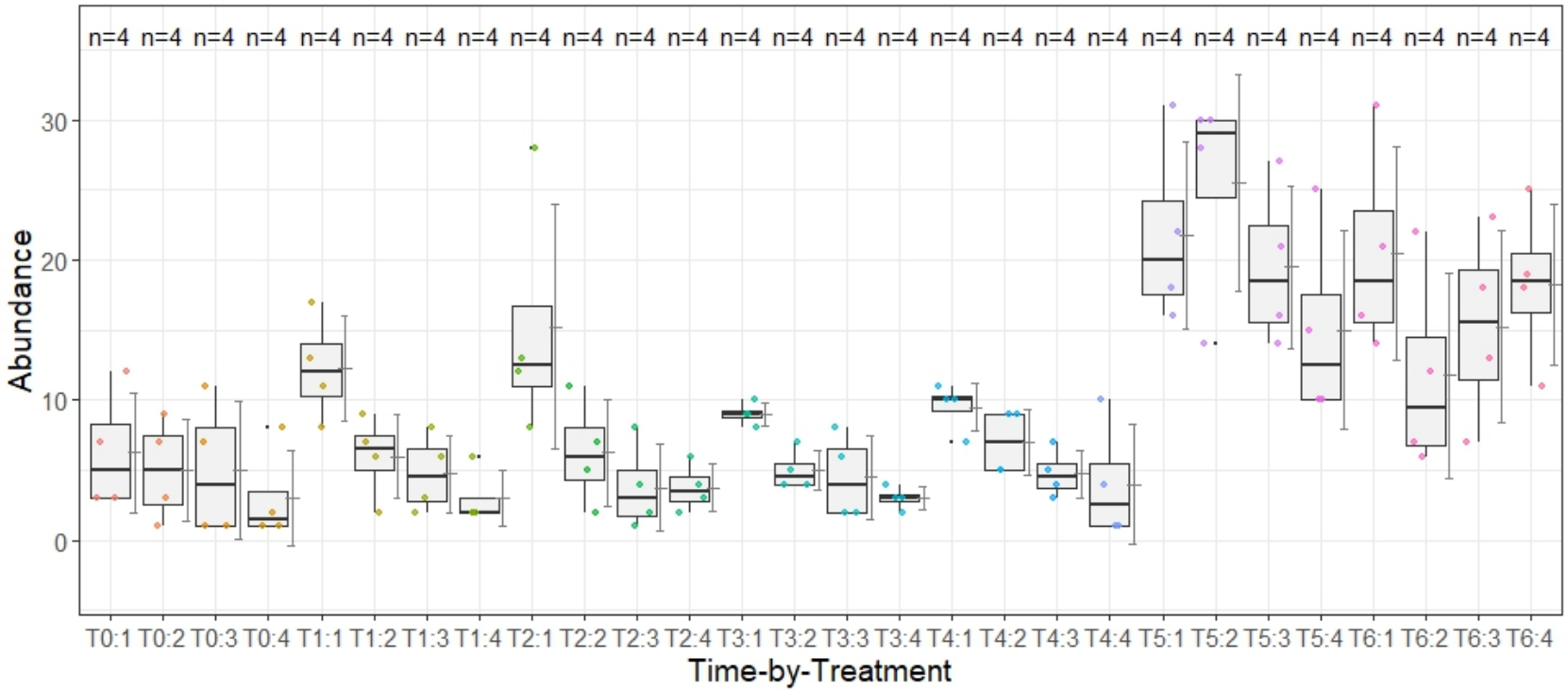
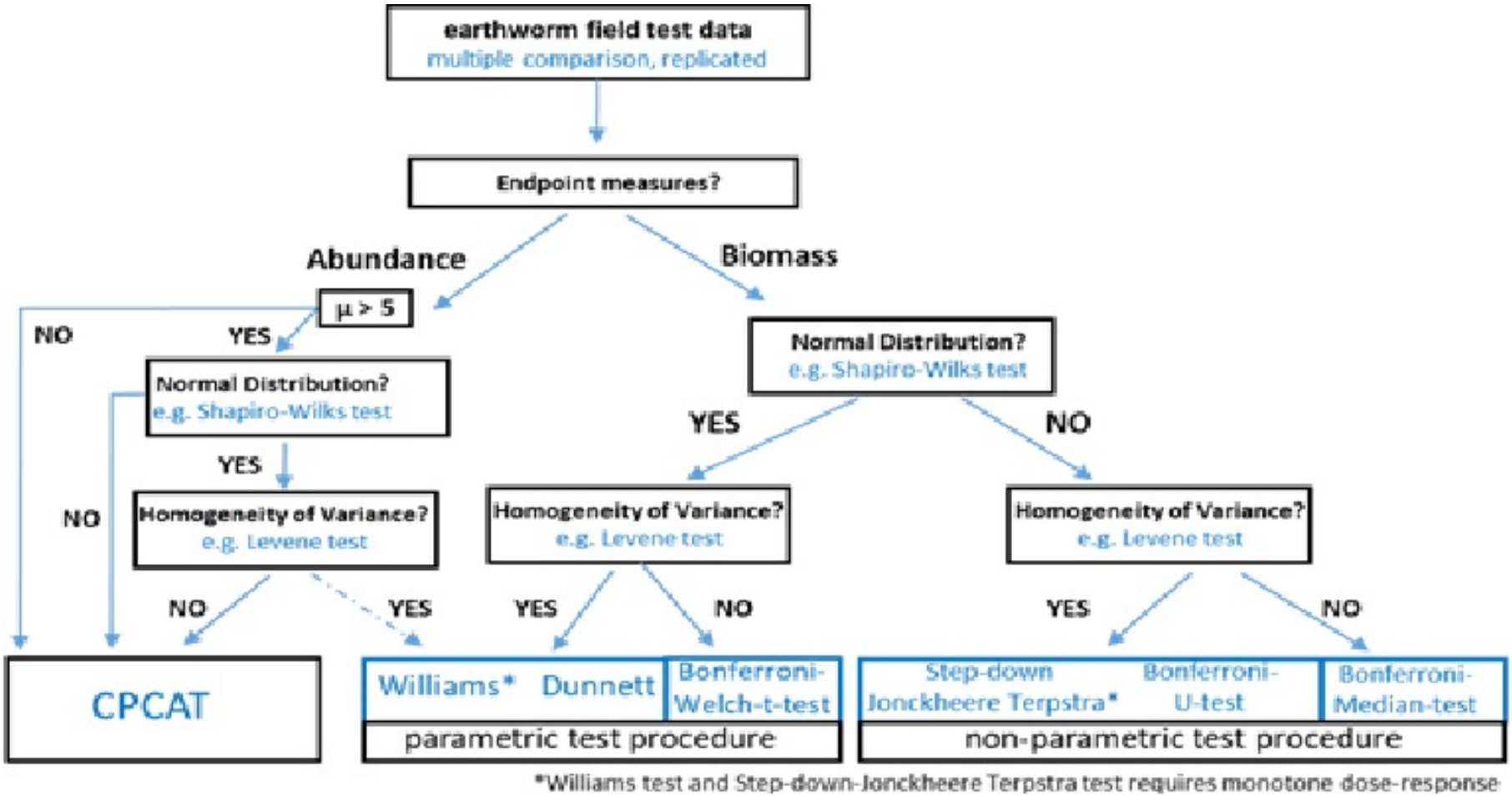


Figure: Ascidea abundance example data

OECD decision tree I

OECD decision tree II



Issues of the OECD decision tree I

- 2 primary endpoints- differently scaled
- Abundances: cutpoint $\mu = 5$ taken from a very old textbook. Hard to defined so simple
- General pre-test/post test dilemma: lack-of-fit tests controlling the less relevant error rate. Equivalence tests needed with a pre-defined tolerance threshold...
- Both test on normality and variance homogeneity too low power for small n_i's
- Alternative test not comparable (e.g. quite different effect sizes)
- **Such decision trees should not be recommended at all**. Alternative: well-chosen, robust tests

Why several concentrations? I

- **Assuming Paracelsus law**: *All things are poison and nothing is without poison. Solely the dose determines that a thing is not a poison*
- OECD propose statistically two quite different approaches: NOEC, BMD
- See Zhenglei's talk on BMD in ecotox
- Both approaches with pros and cons. In the following NOEC only here

Estimate NOEC I

- NOEC depends on the unadjusted effect size δ. Yes, it should. But also on $n_i, s_i, C_k, \Delta C_i, ...$
- OECD **design** recommendation represents a sort of standardization of this point-zero-null-hypotheses tests
- First impulse: ordered concentrations require order restricted tests (to increase power)
- BUT-they use aggregations of C_i and this **biased NOEC estimation**. Williams trend test as a simple example for a specific decreasing plateau-shaped profile:
  $H_1 : \mu_0 = 5 > \mu_1 = 3 < \mu_2 = 4 > \mu_3 = 3.5$:

$\pi_{Du_{1-0}}$	$\pi_{Du_{2-0}}$	$\pi_{Du_{3-0}}$	$\pi_{Wi_{1-0}}$	$\pi_{Wi_{2-0}}$	$\pi_{Wi_{3-0}}$
0.90	0.11	0.66	0.71	**0.71**	0.83

Table: Per-pair powers of Dunnett and Williams test for a simulated specific plateau-shaped alternative

Estimate NOEC II

- Notice, the per-pair power is relevant for NOEC, whereas any-pairs power is commonly published
- Alternative: Dunnett test.
- But Dunnett test may be biased when heterogeneous variances occur- problematic when in the non-NOEC concentrations
- An example- compared with unbiased Welch-type-df modification:

Estimate NOEC III

NOEC	$s_i \Uparrow$	Original Dunnett			Welch-type Dunnett		
		D_{1-0}	D_{2-0}	D_{3-0}	W_{1-0}	W_{2-0}	W_{3-0}
2	-	0.02	0.02	0.89	0.02	0.02	0.85
2	3	0.00	0.00	0.35	0.02	0.02	0.15
2	2	0.00	0.07	**0.23**	0.02	0.02	0.83
2	1	0.07	0.01	**0.24**	0.02	0.01	0.82
2	0	0.07	0.07	0.37	0.03	0.03	0.21
1	-	0.02	0.89	0.88	0.93	0.84	0.83
1	3	0.00	**0.24**	0.38	0.02	0.83	0.14
1	2	0.00	0.36	**0.22**	0.02	0.13	0.83
1	1	0.08	**0.24**	**0.24**	0.02	0.82	0.82
1	0	0.07	0.36	0.37	0.03	0.19	0.20
0	-	0.90	0.91	0.89	0.97	0.85	0.85
0	3	**0.26**	0.24	0.36	0.82	0.84	0.15
0	2	**0.23**	0.36	0.23	0.82	0.14	0.81
0	1	0.38	0.25	0.24	0.15	0.82	0.82
0	0	0.41	0.36	0.35	0.20	0.19	0.20

Table: Simulated per-pair power estimates of Dunnett and Welch-type Dunnett procedure for selected NOEC's and patterns of variance heterogeneity: fair power loss; bias

Estimate NOEC IV

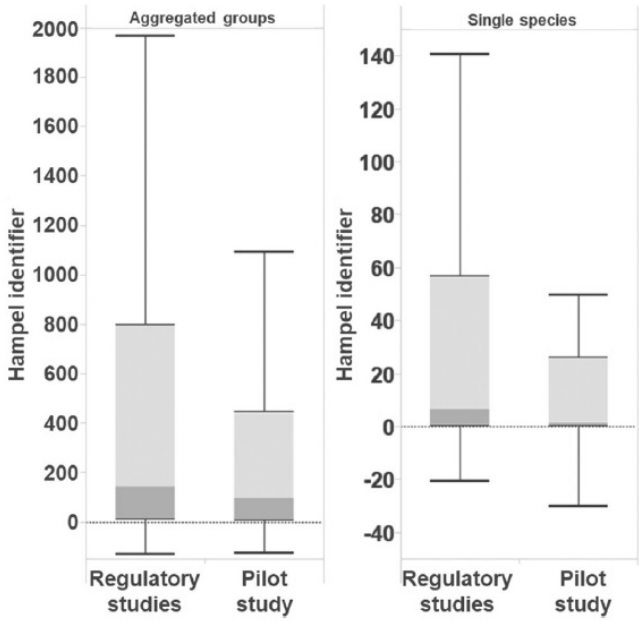
- Recommendation: Estimate NOEC neither by Williams, nor by original Dunnett $\Rightarrow$ use Welch/sandwich modifications!

Count data issues I

- Two primary endpoints per taxonomic level: i) abundance, ii) biomass. 1st count, 2nd continuous
- Empirically heterogeneous variances are observed
- For count data we have TWO related effects:
  - overdispersion for the count variable in itself
  - varying overdispersion with concentration -analogous to heteroscedasticity

Count data issues II

- UBA itself:



- Really challenging: evaluation of overdispersed count data in low n_i and $k + 1$ designs!

OECD proposed test: CPCAT I

- CPCAT based on 2 principles:
  1. CP: Closed testing procedure for comparing C_i's vs. $C-$
  2. CAT: permutative version of LR-test for comparing Poisson distributed counts
- CPCAT's main idea: feasible for small sample sizes count data
- CP-part ok, but does not provide (interpretable) confidence intervals
- CAT part problematic when data overdispersed.... and some data are severe overdispersed

- *OECD: 'The theoretical distribution assumption of earthworm abundance field test data follows a Poisson model'*. **Violates basic stats paradigm.** *'All models are wrong, some are useful' G.Box*

OECD proposed test: CPCAT II

- **Aggregated data:** i) over 4 traps, ii) over taxa.
  - **Stats**: The sum over Poisson variables is only Poisson for complete independence [1]. But they are dependent per definition
  - **Empirical**: UBA data reveal both under and overdispersion, rarely near-to-Poisson data in histor. data [2]
  - **CP-CAT**: *'over-dispersion reduced the statistical power of the CPCAT'* Lehmann et al. 2018.
- **Special features:**
  - **small** n_i
  - **not just overdispersion, but concentration-specific dispersions** similar to variance heterogeneity in Gaussian models

Distrib.	MLT	Nonpar.	CPCAT
Poisson	0.05	0.05	0.04
over	0.05	0.07	0.13
under	0.06	0.05	0.03

OECD proposed test: CPCAT III

- Properties of CPCAT
  1. Falsely small p-values when data are overdispersed- i.e. in the most cases
  2. Falsely large p-values when data are underdispersed- i.e. in some cases
  3. Appropriate p-values when data are exactly Poisson distributed i.e. in rare case

OECD proposed test: CPCAT IV

- Empirical power (Without FWER control(); **max power**)

True NOEC	Distrib.	MLT	Nonpar.	CPCAT
3	Poisson	0.76	**0.82**	**0.82**
3	over	0.58	**0.60**	(0.69)
3	under	0.83	**0.88**	0.67
2	Poisson	0.89	**0.93**	0.91
2	over	0.76	**0.83**	(0.88)
2	under	0.90	**0.94**	0.90
1	Poisson	0.92	0.95	**0.96**
1	over	0.86	**0.94**	(0.96)
1	under	0.94	0.95	**0.97**

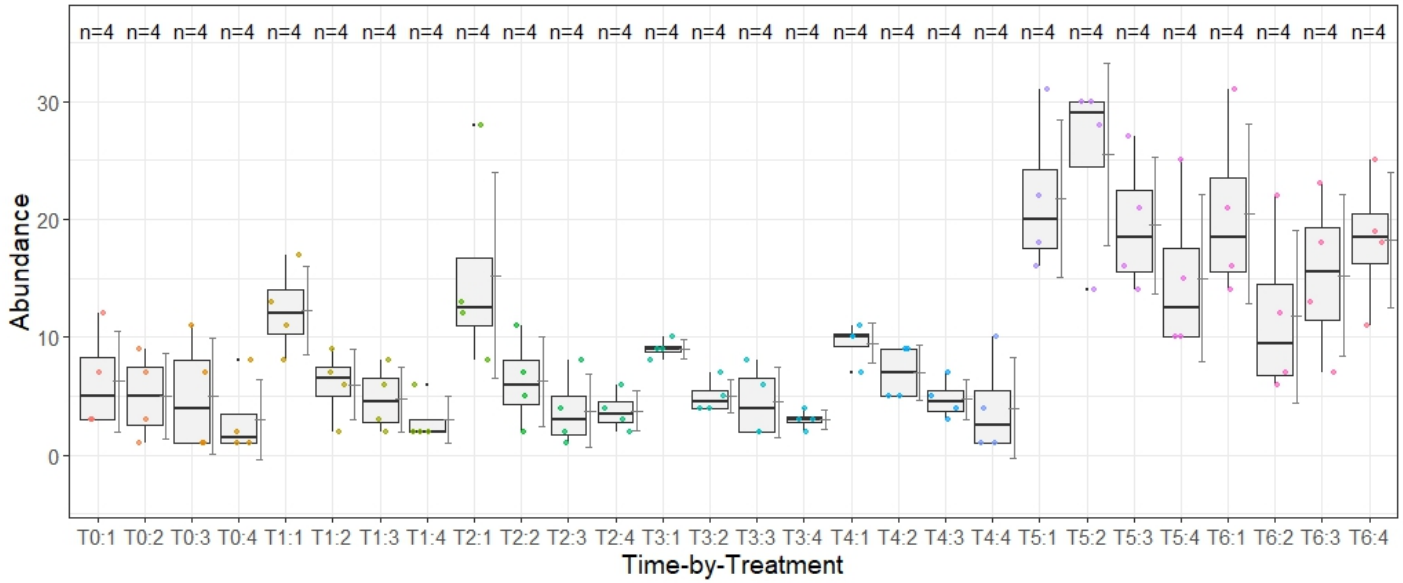
- **My advice: do not use CPCAT for routine analysis**

Alternatives I

- Alternative I: Dunnett test, modified against variance heterogeneity using Welch-type df's [3]
- Alternative II: Nonparametric Dunnett-type test based on global ranks for relative effect sizes [5]. Can be used for both not-rare abundances and biomass
- Alternative III: Dunnett-type test based on most likely transformations sensitive for location/scale/shape effects [4]
- Use simultaneous two-sided $(1 - 2\alpha)$ confidence intervals: i) proof of hazard and safety, ii) decreasing effect at any monitoring time, possible followed by an increase later

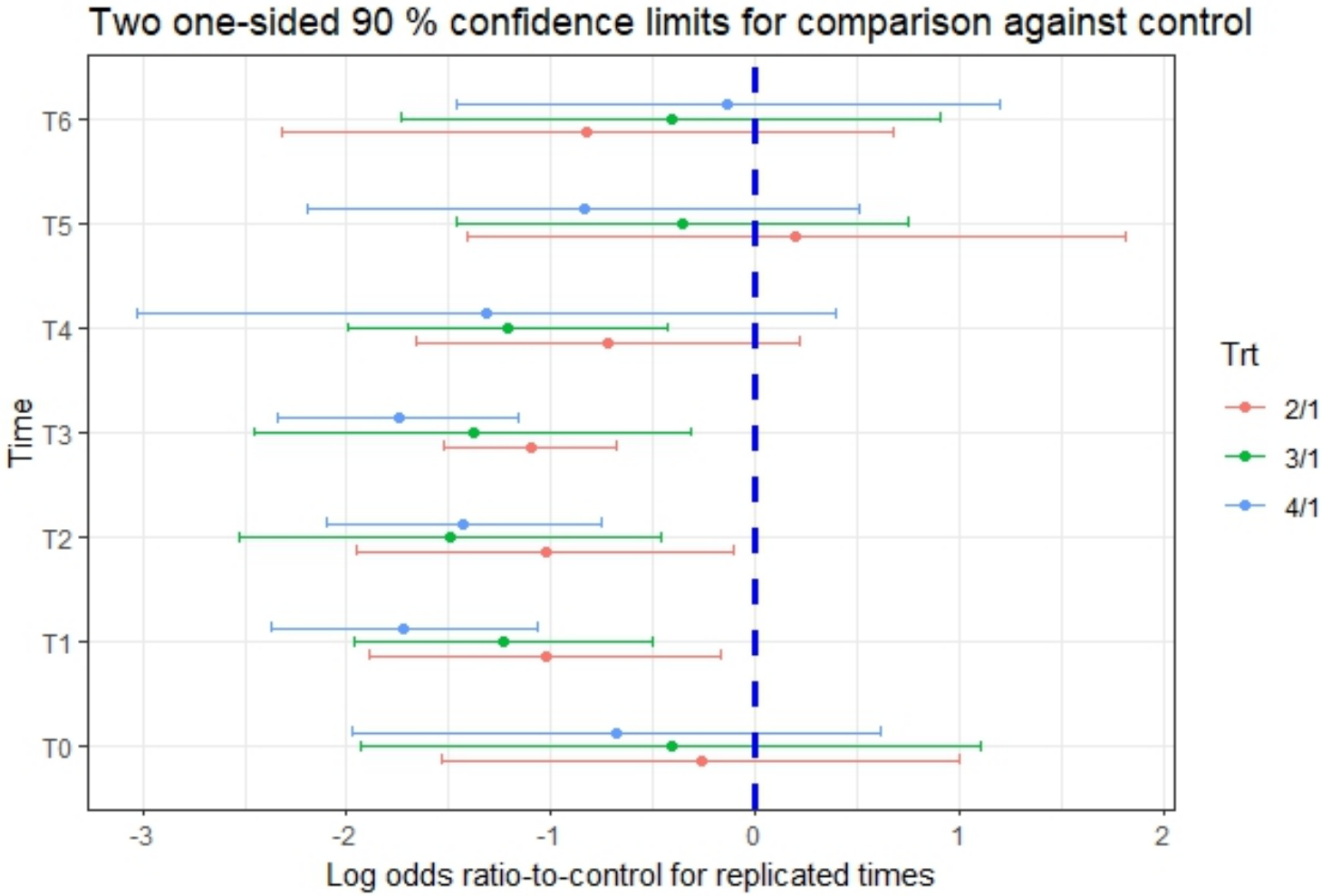
A joint approach: abundance example I

- The abundances of Ascidea in a complete randomized field trial using control and three concentrations (1,2,3,4)



- Nonparametric Dunnett-type test based on global ranks for relative effect sizes [5].
- The confidence intervals for log odds ratios versus control as effect size are presented for each sampling time (including pre-sampling T0):

A joint approach: abundance example II



A joint approach: abundance example III

- At pre-sampling time T0 near-to-equivalence can be concluded, i.e. no serious randomization bias occurred
- Already at T1 a significant decrease of the abundance in each of the concentration occurred, lasting until T3
- Starting at T4 a recovery effect can be observed, which becomes more pronounced at T5, 6.

Take home message I

- Estimate NOEC or BMD. The 1st is data- and design-dependent, the 2nd requires species-specific benchmark thresholds (BMR)
- Neither use CP-CAT nor Dunnett original test (nor any of the OECD proposed tests)
- Use the nonparametric Dunnett-type procedure: robust against variance heterogeneity and overdispersion
- Use confidence limits
- Further issue: KOVAR using pre-sampling data (under work)
- Related R-code available

Appendix: How to model the various sampling times T_t? I

- First impuls: repeated measures analyses by mixed effect model or summarizing approaches (AUC) or even multivariate tests. NO!
- Second: using both T_0 AND T_t -separately
  - T_0 approach I): OECD recommends- demonstrate in each treatment group (before substance administration) a sufficient abundance. Using one-sided CI
  - T_0 approach II): OECD recommends- demonstrate no bias between treatment groups. Dunnett-type equivalence approach. Using 2-sided $(1 - 2\alpha)$ CI's
  - Possible T_0 approach III): KOVAR
- Main objective: demonstrated a possible DECREASING abundance at any sampling time (will be species and ... dependent). Multiplicity-adjusted approaches are possible, but rather conservative (many $t's, n_i = 6$)- performed 1-sided CI unadjusted

Appendix: How to model the various sampling times T_t? II

- Already such a claim for at least ONE species (and/or its aggregations) could be a final outcome of the trial. BUT
- Analyze for a possible recovery, i.e. an following increase of abundance for non-inferiority up to superiority- again by means of an one-sided CI
- Do both together: by two-sided $(1 - 2\alpha)$ CI's
- See the example in a minute

References I

[1] Joel E. Cohen. Sum of a random number of correlated random variables that depend on the number of summands. *Am. Stat.*, 73(1):56–60, 2019.

[2] Benjamin Daniels. Application of the closure principle computational approach test to assess ecotoxicological field studies: Comparative analysis using earthworm field test abundance data. *Environmental Toxicology and Chemistry*, 40(6):1750–1760, 2021.

[3] M. Hasler and L. A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 50(5):793–800, October 2008.

[4] T. Hothorn, L. Most, and P. Buhlmann. Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134, March 2018.

[5] F. Konietschke and L. A. Hothorn. Rank-based multiple test procedures and simult. confidence intervals. *E. J. Stat.*, 6:738–759, 2012.