NCSC2010 Lyon

Williams-type procedures on monotone trend *for normal, non-normal, ordered categorical, proportion and poly-k data, with application in toxicology- using R*

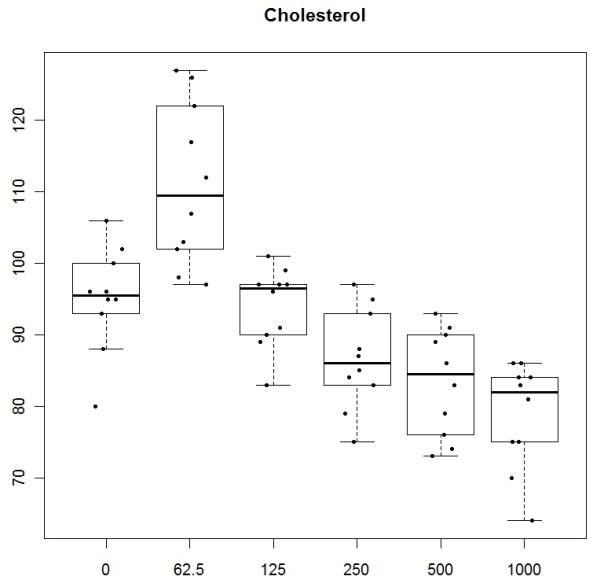
Ludwig A. Hothorn
hothorn@biostat.uni-hannover.de
Institute of Biostatistics, Leibniz University Hannover, Germany

September 21, 2010

Four motivating examples I

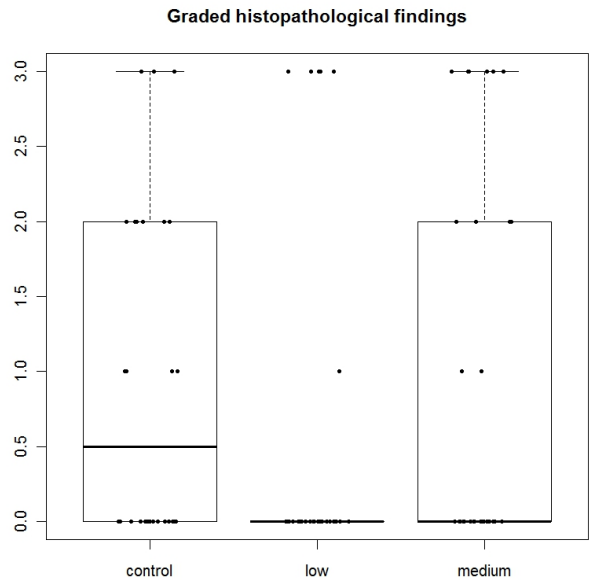
- In the proof of hazard of in-vivo repeated toxicity studies according to OECD guidelines a common randomized design is used: $C, D_{low}, D_{med}, D_{high}$ but different-scaled multiple endpoints occur

- Example I: **Continuous endpoints**: 13 weeks feeding study on Sodium dichromate dihydrate in F344 Rats was downloaded from NTP, e.g. the endpoint cholesterol



Four motivating examples II

- Example II: **Histopathological findings**: incidences of tubular epithelia hyaline droplet degeneration in male rats were reported [WSI+07] for a 28-day oral dose toxicity study of nonylphenol to 0/10, 0/10,3/10,8/10
- Example III: **Graded findings**: non-neoplastic lesions in the P-Cresidine carcinogenicity study on each 30 male mice hyperplasia in parotid gland



Four motivating examples III

- Example IV: **Tumor data in long-term studies** skin fibroma of the NTP study (2000) on the carcinogenic potential of methyleugenol

dose	0 mg/kg	37 mg/kg	75 mg/kg	150 mg/kg
Crude Rate	1/50	9/50	8/50	5/50
Crude Percent	2%	18%	16%	10%

Table: Chronic toxicity study on methyleugenol

- Common evaluation according to NTP: comparisons versus control without order restriction, e.g. Dunnett or Dunn procedure, or with order restriction, e.g. Williams or Shirley procedure

Trend tests and related simultaneous confidence intervals I

- Important criteria of relevance in the proof of hazard: **a significant trend**.
- Trends with a-priori unknown shapes can be tested by the covariate DOSE, whereas the problem of model selection occur- see tommorow's talk by Ch. Ritz or by the factor dose, assuming an appropriate dose metrics, e.g. the common log-scale. Pro's and con's of both approaches, here the multiple contrast test
- Preferring simultaneous confidence intervals:
  - i interpretation together with an effect size,
  - ii one-sided sCI allows both proof of hazard AND proof of safety(not today)
- Question: what means trend? Two criteria:
  - i one-sided
  - ii monotone $H_1 : \mu_C \leq \mu_1 \leq ... \leq \mu_k$ i.e. all possible elementary hypotheses, not just a linear trend.

Trend tests and related simultaneous confidence intervals II

- Therefore, a trend test must be sensitive against all possible elementary alternatives, not against just one, e.g. the linear as the wide-spread used Cochran-Armitage trend test [Arm55] for proportions or the Jonckheere trend test for pairwise ranks.
- At least two approaches:
  i. MLE-test acc. to [Bar59] **quadratic test statistics**
  ii. MCT **linear test statistics**
- A trend test, which compares vs. control: Williams trend test [Wil71].
- A contrast is a suitable linear combination of means: $\sum_{i=0}^{k} c_i \bar{x}_i$.
  A contrast test is standardized $t_{Contrast} = \sum_{i=0}^{k} c_i \bar{x}_i / S \sqrt{\sum_{i}^{k} c_i^2 / n_i}$
  where $\sum_{i=0}^{k} c_i = 0$ guaranteed a $t_{df,1-\alpha}$ distributed level-α-test.

Trend tests and related simultaneous confidence intervals III

- A multiple contrast test is defined as maximum test: $t_{MCT} = max(t_1, ..., t_q)$ which follows jointly $(t_1, \ldots, t_q)'$ a q-variate t- distribution with degree of freedom df and the correlation matrix R, with $\rho_{ab} = \frac{\sum_{i=1}^{k} a_i b_i / n_i}{\sqrt{\sum_{i=1}^{k} a_i^2 / n_i \sum_{i=1}^{k} b_i^2 / n_i}}$
- **Notice**: With increasing average correlation and lower number of contrasts q the q-variate t-distribution tends to the univariate t-distribution, i.e. the degree of adjustment reduces
- The contrast structure of Williams (1971) procedure [Wil71]

c_i	C	D_1	D_2
c_a	-1	0	1
c_b	-1	1/2	1/2

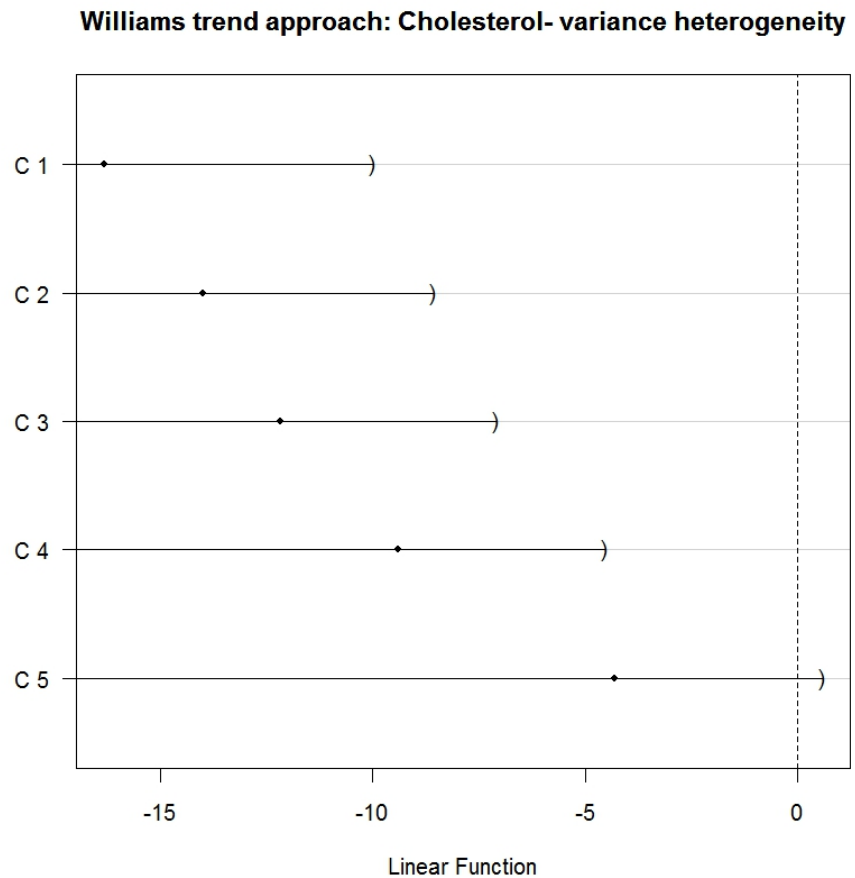
- One-sided (lower) confidence intervals: $[\sum_{i=0}^{k} c_i \bar{x}_i - St_{q,df,R,2-sided,1-\alpha} \sqrt{\sum_i^k c_i^2 / n_i}]$

Trend tests and related simultaneous confidence intervals IV

- Notice: multiplicity-adjusted p-values are available alternatively to simultaneous confidence intervals. And they are compatible
- Variance heterogeneity is more likely in real toxicological data than variance homogeneity, since a possible proportionality between variance and mean
- Three approaches:
  i. Using a sandwich estimator for variance-covariance matrix in the linear model [HSH10]
  ii. Welch-type df-adjustment for multiple contrast tests [Has09],
  iii. Behrens-Fisher modification of non-parametric tests [FK09].
  R-programs are available.

Trend tests and related simultaneous confidence intervals V

- Evaluation of the cholesterol example by means of `multcomp`



Multiple comparisons for ratios of μ_i I

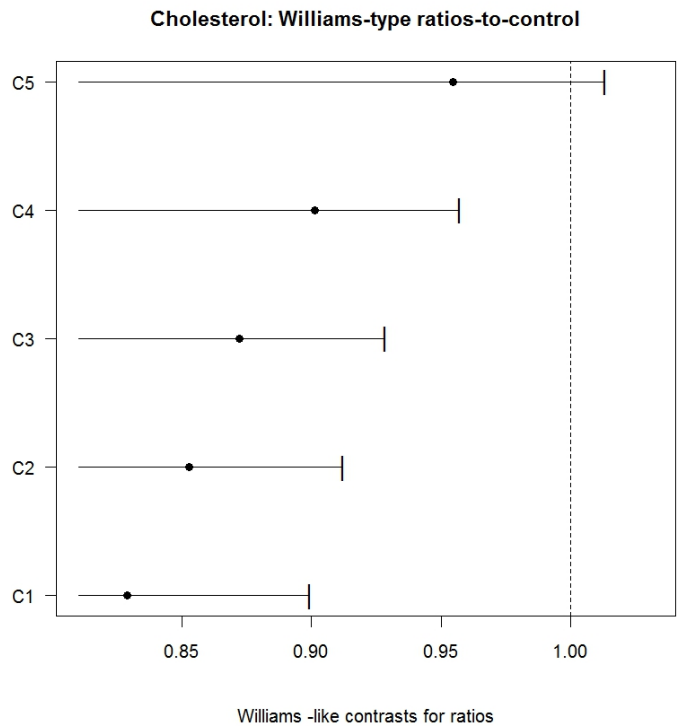
- **Aim**: simultaneous confidence intervals for μ_i/μ_0
- Trick: Re-formulation the ratios in a linear form $Z_{i0} = \bar{x}_i - \theta\bar{x}_0$ [Fie54]
- Simultaneous confidence intervals for the ratios $\gamma_{i0} = \mu_i/\mu_0$

$$\left\{(\widehat{\gamma}_i - G) \pm \left[(\widehat{\gamma}_i - G)^2 - (1-G)\left(\widehat{\gamma}_i^2 - \frac{N}{n_i}G\right)\right]^{\frac{1}{2}}\right\}/(1-G)$$

  $i = 1, \ldots, q$, where $G = S^2 q^2_{\alpha,m,\nu,R}/(N\bar{x}_0^2)$
- Notice, the equi-coordinate percentage point $t_{q,\nu,\mathbb{R},1-\alpha}$ depends on the unknown ratios γ_{i0} by the correlation matrix. Solution: Plug-in [DBGH04] realized in the R package `mratios` [DSH07]

Multiple comparisons for ratios of μ_i II

- Advantage: interpretability of different-scaled multiple endpoints by percentage change
- Evaluation of the cholesterol example by means of `mratios`



Non-parametric approaches and related simultaneous confidence intervals I

- For non-normal data, the trend test according to Shirley [SHI77] is widely used toxicology. I.e. the observations are jointly ranked and Williams' test is applied.
- $H_0^F : F_0 = ... = F_k$ formulated in terms of the distribution functions against the ordered alternative $H_1^F : F_0 \leq \ldots \leq F_k$ with at least one strict inequality $F_i < F_s, i \neq s$. It controls the FWER strongly.
- The distribution of the rank means is unknown under the alternative, neither simultaneous confidence intervals are numerically available for a general unbalanced design, nor power can be estimated.
- Tied or ordered categorical data, such as severity counts, should be analyzed as well.
- Variance heterogeneity occurs frequently; therefore a Behrens-Fisher (BF) version is needed

Non-parametric approaches and related simultaneous confidence intervals II

- Using relative effect size [BM00],[RA08]:

$$p_{01} = \int F_0 dF_1 = P(X_{01} < X_{11}) + 0.5P(X_{01} = X_{11}). \qquad (1)$$

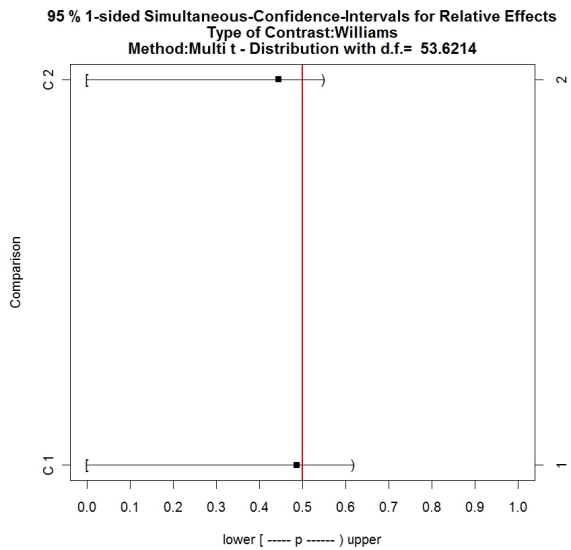
  Hereby, add $0.5P(X_{01} = X_{11})$ for ties

- Relative Shirley-type effects. Treatment effects can be defined by using the relative effect between the distribution of the negative control group F_0 and the distribution of the samples M_ℓ, $\ell = 1, \ldots, q$:

$$\begin{aligned}
p_1 &= p_{0k} \\
p_2 &= \frac{n_{k-1}}{n_{k-1} + n_k} p_{0(k-1)} + \frac{n_k}{n_{k-1} + n_k} p_{0k} \\
&\vdots \\
p_q &= \frac{n_1}{n_1 + \ldots + n_k} p_{01} + \ldots + \frac{n_k}{n_1 + \ldots + n_k} p_{0k}.
\end{aligned}$$

Non-parametric approaches and related simultaneous confidence intervals III

- Shirley-type test for graded histopathological findings by means of `nparcomp`
- Scores data are particularly suitable for statistics of relative effects [RA08] The graded findings [none, Mild, Moderate, Marked] will be transferred into the equal-distant scores [0,1,2,3]



Simultaneous confidence intervals for proportions I

- Rates are rather typically in toxicological studies, e.g. histopathological findings, mortality, tumor rates
- **General contradiction in toxicological risk assessment:** the evaluation of continuous endpoints is powerful and related statistical approaches are widely available- however their predictive value is limited, such as body weight. But, the predictive value of proportions, such as histopathological findings, is larger, but the power is much lower and appropriate statistical approaches are rarely available for such small sample sizes
- Moreover, for sample sizes of $n_i = 50....10$ there is no hope for valid $(1 - \alpha)$ Wald intervals- therefore we need confidence intervals where its coverage probability is also for smaller samples (not really small samples) is approximately 95%
- And, for all proportions a one-sided alternative for an increase is appropriate, never a two-sided alternative

Simultaneous confidence intervals for proportions II

- As effect size the difference of proportions is common (alternatively relative risk)
- One-sided, lower $(1-\alpha)$ Wald-type confidence limits for the difference of the proportions of treatment against those from a control are

$$\left[\sum_{i=1}^{I} c_i p_i - z_{q,R,1-\alpha}\sqrt{\sum_{i=1}^{I} c_i^2 \hat{V}\left(p_i\right);}\right] \tag{2}$$

with $\hat{V}\left(p_i\right) = p_i\left(1-p_i\right)/n_i$ and $z_{q,R,1-\alpha}$ denoting the $(1-\alpha)$ quantile of the q-variate normal distribution whereas its correlation matrix $\mathbf{R}$ depends not only on the known contrast coefficients c_{im} and sample sizes n_i but also on the unknown π_i and $V(p_i)$ where the plug-in of the ML-estimators $\hat{\pi}_i$ and $\hat{V}(\pi_i)$ works well.

- However, Wald limits for binomial proportions are known to keep the $(1-\alpha)$ coverage probability only for asymptotically large sample sizes [AC00], [PB04]

Simultaneous confidence intervals for proportions III

- [AC98] showed that adding a total of four pseudo-observations to the observed successes and failures yields approximate confidence intervals for one binomial proportion with good small sample performance
- One-sided limits was investigated only by [Cai05] in the case of a single binomial proportion, and recently [SV09]

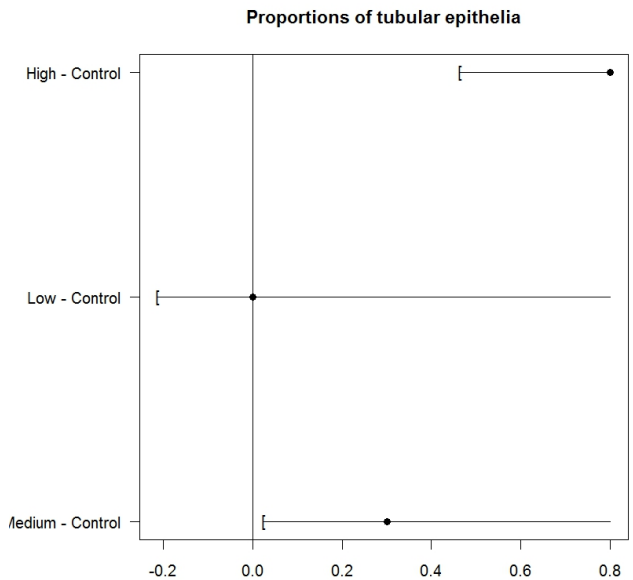
$$\left[\sum_{i=1}^{I} c_i \tilde{p}_i - z_{q,R,1-\alpha}\sqrt{\sum_{i=1}^{I} c_i^2 \tilde{V}\left(\tilde{p}_i\right)}\right] \tag{3}$$

Table: Choices for $\tilde{p}_i$ and $\tilde{V}\left(p_i\right)$

Notation	$\tilde{p}_i$	$\tilde{V}(p_i)$
Wald	Y_i/n_i	$p_i\left(1-p_i\right)/n_i$
add-1	$\left(Y_i+0.5\right)/\left(n_i+1\right)$	$\tilde{p}_i\left(1-\tilde{p}_i\right)/\left(n_i+1\right)$
add-2	$\left(Y_i+1\right)/\left(n_i+2\right)$	$\tilde{p}_i\left(1-\tilde{p}_i\right)/\left(n_i+2\right)$

Simultaneous confidence intervals for proportions IV

- A simulation study [SSH08] indicates the use of the add1 approximation for one-sided lower limits when sample sizes are not too small
- Simultaneous confidence limits for tubular epithelia hyaline droplet degeneration in male rats by means of `MCPAN`.



A Williams-type for overdispersed counts I

Sorry no time today, solved by Gerhard's (2010) recent thesis

A Williams-type for poly-3 estimates I

- In long-term carcinogenicity studies the compound may not only affect the tumor rate but also the mortality in the treatment groups. Two approaches with/without cause-of-death information - here poly-k trend test [BP88] only
- To account for censoring due to treatment-specific mortality,[BP88] proposed the poly-3 adjustment by individual weights $w_{ij} = (t_{ij}/t_{max})^k$.
- These weights result in an adjusted group sample size $n_i^* = \sum_{j=1}^{n_i} w_{ij}$ and therefore in adjusted proportions $p_i^* = y_i/n_i^*$.
- [SSH08] demonstrated there plug-in instead of the crude proportions into Dunnett/Williams procedure

A Williams-type for poly-3 estimates II

- Evaluation of the example: lower 95% Add-1 confidence limits to detected an increasing trend in mortality-adjusted tumor rates

dose	0 mg/kg	37 mg/kg	75 mg/kg	150 mg/kg
Crude Rate	1/50	9/50	8/50	5/50
Crude Percent	2%	18%	16%	10%
Poly-3 adjusted-Rate	1/41.4	9/40.3	8/38.7	5/32.7
Poly-3 adjusted-Percent	0.02%	0.22%	0.21%	0.15%

Table: Chronic toxicity study on methyleugenol

Comparison	estimate	lower limit	adjusted p-value
high vs. control	0.1288	−0.009	0.066
high, medium vs. control	0.1555	0.048	0.005
high, medium, low vs. control	0.1701	0.075	0.0005

Table: Evaluation of the methyleugenol data set using the Williams contrast

Take home message I

- Williams-type MCTs can be recommend for proof of hazard evaluation with control of FWER in toxicology: parametric (difference and ratio), non-parametric and proportion, zero-inflated counts and poly-3 estimates
- For all occuring types of endpoints in a repeated toxicological study **unique Dunnett/Williams approaches** are available
- Related R-libraries available
- Interpretability of sCI should be THE argument of its use instead of p-values
- Now, theses appropriate approaches should used and understood by toxicologists, to avoid sentences like *statistically significant, but biologically not relevant* in future.
- Still open problems, e.g. in the mixed model (work in progress)
- Can be similarly used in clinical dose findings studies, microarray data and ....

References I

[AC98] Agresti, A. ; Coull, B. A.: Approximate is better than "exact" for interval estimation of binomial proportions. In: *American Statistician* 52 (1998), Mai, Nr. 2, S. 119–126

[AC00] Agresti, A. ; Caffo, B.: Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. In: *American Statistician* 54 (2000), November, Nr. 4, S. 280–288

[Arm55] Armitage, P: Tests for Linear Trends in Proportions and Frequencies. In: *Biometrics* 11 (1955), Nr. 3, S. 375–386

[Bar59] Bartholomew, D. J.: A Test Of Homogeneity For Ordered Alternatives .2. In: *Biometrika* 46 (1959), Nr. 3-4, S. 328–335

[BM00] Brunner, E. ; Munzel, U.: The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. In: *Biometrical Journal* 42 (2000), Nr. 1, S. 17–25

[BP88] Bailer, A.J. ; Portier, C. J.: Effects Of Treatment-Induced Mortality And Tumor-Induced Mortality On Tests For Carcinogenicity In Small Samples. In: *Biometrics* 44 (1988), Juni, Nr. 2, S. 417–431

[Cai05] Cai, T. T.: One-sided confidence intervals in discrete distributions. In: *Journal Of Statistical Planning And Inference* 131 (2005), April, Nr. 1, S. 63–88

[DBGH04] Dilba, G. ; Bretz, E. ; Guiard, V. ; Hothorn, L. A.: Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. In: *Methods Of Information In Medicine* 43 (2004), Nr. 5, S. 465–469

[DSH07] Dilba, G. ; Schaarschmidt, F. ; Hothorn, L.A.: Inferences for ratios of normal means. In: *R News* 7 (2007), S. 20–23

[Fie54] Fieller, E. C.: Some Problems In Interval Estimation. In: *Journal Of The Royal Statistical Society Series B-Statistical Methodology* 16 (1954), Nr. 2, S. 175–185

[FK09] F. Konietschke, L. A. H. E. Brunner B. E. Brunner: Nonparametric relative contrast effects: Asymptotic Theory and Small Sample Approximations / University of Goettingen and Leibniz University of Hannover. 2009. – Forschungsbericht

[Has09] Hasler, M.: *Extensions of Multiple Contrast Tests*, Gottfried Wilhelm Leibniz Universität Hannover, Diss., 2009

[HSH10] Herberich, E. ; Sikorski, J. ; Hothorn, T.: A Robust Procedure for Comparing Multiple Means under Heteroscedasticity in Unbalanced Designs. In: *Plos One* 5 (2010), März, Nr. 3, S. e9788

References II

[PB04] Price, R. M. ; Bonett, D. G.: An improved confidence interval for a linear function of binomial proportions. In: *Computational Statistics & Data Analysis* 45 (2004), April, Nr. 3, S. 449–456

[RA08] Ryu, E. J. ; Agresti, A.: Modeling and inference for an ordinal effect size measure. In: *Statistics In Medicine* 27 (2008), Mai, Nr. 10, S. 1703–1717

[SHI77] SHIRLEY, E: NONPARAMETRIC EQUIVALENT OF WILLIAMS TEST FOR CONTRASTING INCREASING DOSE LEVELS OF A TREATMENT. In: *BIOMETRICS* 33 (1977), Nr. 2, S. 386–389. – ISSN 0006–341X

[SSH08] Schaarschmidt, F. ; Sill, M. ; Hothorn, L. A.: Poly-k-trend tests for survival adjusted analysis of tumor rates formulated as approximate multiple contrast test. In: *Journal Of Biopharmaceutical Statistics* 18 (2008), Nr. 5, S. 934–948

[SV09] Schaarschmidt, F. ; Vaas, L.: Analysis of Trials with Complex Treatment Structure Using Multiple Contrast Tests. In: *Hortscience* 44 (2009), Februar, Nr. 1, S. 188–195

[Wil71] Williams, D. A.: Test For Differences Between Treatment Means When Several Dose Levels Are Compared With A Zero Dose Control. In: *Biometrics* 27 (1971), Nr. 1, S. 103–&

[WSI+07] Woo, G. H. ; Shibutani, M. ; Ichiki, T. ; Hamamura, M. ; Lee, K. Y. ; Inoue, K. ; Hirose, M.: A repeated 28-day oral dose toxicity study of nonylphenol in rats, based on the 'Enhanced OECD Test Guideline 407' for screening of endocrine-disrupting chemicals. In: *Archives Of Toxicology* 81 (2007), Februar, Nr. 2, S. 77–88