Statistics in Toxicology Using R

Ludwig A. Hothorn

Institut für Biostatistik, Leibniz Universität

Hannover, Germany

CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the

Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
The R Series

Series Editors

John M. Chambers
Department of Statistics
Stanford University
Stanford, California, USA

Torsten Hothorn
Division of Biostatistics
University of Zurich
Switzerland

Duncan Temple Lang
Department of Statistics
University of California, Davis
Davis, California, USA

Hadley Wickham
RStudio
Boston, Massachusetts, USA

Aims and Scope

This book series reflects the recent rapid growth in the development and application of R, the programming language and software environment for statistical computing and graphics. R is now widely used in academic research, education, and industry. It is constantly growing, with new versions of the core software released regularly and more than 7,000 packages available. It is difficult for the documentation to keep pace with the expansion of the software, and this vital book series provides a forum for the publication of books covering many aspects of the development and application of R.

The scope of the series is wide, covering three main threads:

- Applications of R to specific disciplines such as biology, epidemiology, genetics, engineering, finance, and the social sciences.
- Using R for the study of topics of statistical methodology, such as linear and mixed modeling, time series, Bayesian methods, and missing data.
- The development of R, including programming, building packages, and graphics.

The books will appeal to programmers and developers of R software, as well as applied statisticians and data analysts in many fields. The books will feature detailed worked examples and R code fully integrated into the text, ensuring their usefulness to researchers, practitioners and students.

Published Titles

Stated Preference Methods Using R, *Hideo Aizaki, Tomoaki Nakatani, and Kazuo Sato*

Using R for Numerical Analysis in Science and Engineering, *Victor A. Bloomfield*

Event History Analysis with R, *Göran Broström*

Computational Actuarial Science with R, *Arthur Charpentier*

Statistical Computing in C++ and R, *Randall L. Eubank and Ana Kupresanin*

Basics of Matrix Algebra for Statistics with R, *Nick Fieller*

Reproducible Research with R and RStudio, Second Edition, *Christopher Gandrud*

R and MATLAB® *David E. Hiebeler*

Statistics in Toxicology Using R *Ludwig A. Hothorn*

Nonparametric Statistical Methods Using R, *John Kloke and Joseph McKean*

Displaying Time Series, Spatial, and Space-Time Data with R, *Oscar Perpiñán Lamigueiro*

Programming Graphical User Interfaces with R, *Michael F. Lawrence and John Verzani*

Analyzing Sensory Data with R, *Sébastien Lê and Theirry Worch*

Parallel Computing for Data Science: With Examples in R, C++ and CUDA, *Norman Matloff*

Analyzing Baseball Data with R, *Max Marchi and Jim Albert*

Growth Curve Analysis and Visualization Using R, *Daniel Mirman*

R Graphics, Second Edition, *Paul Murrell*

Introductory Fisheries Analyses with R, *Derek H. Ogle*

Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving, *Deborah Nolan and Duncan Temple Lang*

Multiple Factor Analysis by Example Using R, *Jérôme Pagès*

Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R, *Daniel S. Putler and Robert E. Krider*

Implementing Reproducible Research, *Victoria Stodden, Friedrich Leisch, and Roger D. Peng*

Graphical Data Analysis with R, *Antony Unwin*

Using R for Introductory Statistics, Second Edition, *John Verzani*

Advanced R, *Hadley Wickham*

Dynamic Documents with R and knitr, Second Edition, *Yihui Xie*

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20151116

International Standard Book Number-13: 978-1-4987-0128-0 (eBook - PDF)

Contents

List of Figures

List of Tables

Preface

When you write a book on *statistics in toxicology* you are sitting on the fence: For statisticians much is known (or at least not exciting), but a large number of toxicologists will be overwhelmed. This book was prepared with a minimum of mathematical formulas, but on the basis of real data examples that are explicitly evaluated with specific R programs. Hence, the second part of the title "*Using R*" was selected.

The book is primarily structured according to selected toxicological assays, and, based on statistical methods. This style comes at the price of somewhat annoying repetitions and many cross-references.

The book is available in print. However, in addition there is a `knitr` [409] document and the package "*SiTuR*", which on the notebook under R can run all data examples by definition. This provides an additional way of learning: Replacing selected data examples through your own data.

Looking at current toxicological publications closely, notice that the core problem *is a significant test of a criterion for hazard*—or perhaps more important: *is non-significance a criterion for harmlessness*—is not really solved. The apparent contradiction between statistical significance and biological relevance has greatly reduced the importance of statistics in decision making—thus reducing the value of statistical methods as a whole in toxicology. One aim of this book is to highlight ways out of this dilemma.

Following the establishment of regulatory toxicology in the 1970s, many statistical methods for toxicology have been published. Regardless, the recommendations for statistical analysis remain remarkably imprecise in most toxicological guidelines; even the pinnacle, the experimental design, remains imprecise. The second aim of this book is to present assay-specific proposals, e.g., for the *in vitro* micronucleus assay. It provides suggestions, not recommendations, and is certainly not a cookbook!

The statistical complexity varies substantially in this book: From the t-test to the mixing distribution approach in the Comet assay trying to describe the appropriate evaluation.

The statistical analysis of experimental data is a current standard in toxicology. In almost every paper, statistical methods are used. Tests and their p-values are dominating today—explicitly or at least implicitly for claiming significant effects. Toxicology is a broad field. This book focuses on standardized bioassays for chemicals (by OECD guidelines, e.g., [283]), drugs (by ICH guidelines, e.g., [199]) and environmental pollutants (by EPA, OECD and ECOTEC guidelines [12, 282]) and consequently hypothesis testing is the focus. Accordingly, the book is organized initially by selected bioassays: i) short-term repeated toxicity studies, ii) long-term carcinogenicity assays, iii) studies on reproductive toxicity, iv) mutagenicity assays, and v) toxicokinetic studies. Methodically oriented chapters follow: vi) proof of safety, vii) toxicogenomics, viii) toxicokinetics, ix) analysis of interlaboratory studies and x) modeling of dose–response relationships for risk assessment.

For some readers, this text will appear testing heavy. Yes, it is. Six arguments speak in favor of hypothesis testing against modeling: i) the only focus in the related US National

Toxicology Program (NTP) recommendation [9], ii) the main focus of most guidelines, iii) commonly used designs with two or three doses only (+ zero dose control), iv) availability of dose metameters only–probably not simply related to the concentrations at the target in the bioassays nor to the human exposure, v) superiority of modeling only when very specific assumptions hold true and vi) the preferred method in recent publications in experimental toxicology.

The challenge was a representation style using few formulas for sometimes complex statistical methods. The first focus is on raw *data*, *data* structures, and *data* models. The second focus is **using R packages**. The third focus is reproducible publication [191] by means of a single `knitr` document containing all the raw data, R code, text, figures, tables, statistical outcomes, references. That is, educationally the following order is adhered to: a brief explanation of the toxicological problem, followed by the equally short explanation of the statistics, the matching data example (structure, raw data, visualization), the R code, the outcomes and their interpretation.

This allows a toxicologist to select a certain bioassay, e.g., the Ames assay, to understand the specific data structure (counts, small samples, non-monotone dose-response effect etc.), to run the R code with the data example, understand the test outcome and the interpretation, and replace the data set with his/her own data and run again.

All data that are used in this book are either available in the package `SiTuR` or otherwise in cited R packages.

To the reader I want to apologize for the less than perfect figures and tables; my focus was direct processability in `knitr`, i.e., to achieve a reproducible document.

Acknowledgments

Without the cooperation of the following colleagues, this book would not have been possible: G. Dilba Djira (Brookings), G. Gerhard (Christchurch), M. Hasler (Kiel), E. Herberich (Ingelheim), T. Hothorn (Zurich), Th. Jaki (Lancaster), A. Kitsche (Hannover), F. Konietschke (Dallas), R. Kuiper (Utrecht), P. Pallmann (Lancaster), Ch. Ritz (Copenhagen), F. Schaarschmidt (Hannover), and Z. Zhkedy (Hassel) —this is particularly true for making available the R packages `multcomp, mratios, pairwiseCI, MCPAN, ETC, goric, drc, bmd, statint, toxbox, Isogene, coin, nparcomp, PK`. Also, I would like to acknowledge the free availability of numerous raw data by the US NTP database (`http://ntp.niehs.nih.gov/results/dbsearch/index.html`).

Moreover, I thank L. Edler (Heidelberg), A. Kopp-Schneider (Heidelberg), R. Pirow (Berlin), C. Vogel (Hannover) and A. Krueger (Braunschweig) for the critical comments when reading draft chapters.

I am grateful to three anonymous reviewers for the constructive and encouraging comments which improved the quality substantially.

1

Principles

1.1 Evaluation of short-term repeated toxicity studies

The principles of design and evaluation are discussed in the following by means of selected items in short-term repeated toxicity studies. The aim of repeated dose toxicity studies is to characterize the toxicological profile of a compound usually administered between 4 weeks [283] and 3 months [284]. Although baseline values or data of a recovery period are available in some studies, the typical data are multiple endpoints at the end of the administration period, such as continuous endpoints (e.g., hemoglobin), rates (e.g., proportions of histopathological findings), or ordered categorical data (e.g., graded histopathological findings). The standard design uses a negative control (C) and several doses $D_1, ..., D_k$ where $k = 3$ is common. A one-way layout will be considered during evaluation, i.e., both sexes are analyzed independently (see their joint analysis in Section 2.4.1). The comparisons of doses versus control are performed by unadjusted two-sample tests (see Section 1.3), or the Dunnett procedure [102] without order restriction (see Section 2.1.1) or the Williams procedure [400] with order restriction (see Section 2.1.2). The usual sample sizes per group i, $n_i = 10$ in rodent studies allow the use of standard tests, including asymptotic ones, though only under particular assumptions, or only for selected endpoints. For all other endpoints and particularly small sample sizes like in dog studies, specific tests are needed. This small sample size restriction is the first feature of statistics in toxicology; it is repeatedly discussed in the book.

Features:

- Principle of statistical inference in toxicology
- Proof of hazard without control of the familywise error rate (FWER)
- Two sample comparisons for continuous, skewed variables, proportions and counts

1.2 Selected statistical problems

1.2.1 Data visualization by barcharts or boxplots?

Up to now the most common style of data presentation are tables containing group-specific means, standard deviations, and sample sizes, commonly for multiple endpoints; see e.g.,

[376] in Figure 1.1.

Table 1 Effect of tBOOH on cellular integrity, redox status, and Ca^{2+} levels

	DMSO[a]	*t*BOOH[b]
Cellular viability (% of total cells)	96 ± 3	62 ± 3*
LDH release (% of total cell content)	32 ± 2	42 ± 3*
ATP cellular content (μmol/10^6 cells)	48 ± 6	43 ± 3
MDA cellular content/nmol/mg prot.)	1.3 ± 0.2	8.0 ± 0.3*
Total glutathione (mg/mg of protein)	27 ± 2	17 ± 1*
GSSG-to-total glutathione ratio (%)	3.1 ± 0.1	4.1 ± 0.2*
Cytosolic free Ca^{2+} (nM)	119 ± 11	3029 ± 474*

Cells were pretreated for 15 min either with *t*BOOH (500 μM) or with DMSO (controls)

LDH lactate dehydrogenase, *ATP* adenosine triphosphate, *MDA* malondialdehyde, *GSSG* oxidized glutathione

* $p < 0.05$ versus DMSO, for $n = 4$–10

FIGURE 1.1: Example for data summary table.

Bar charts are commonly used as well, e.g., [98] used barcharts (including SEM) and letters of significance (Figure 1.2).

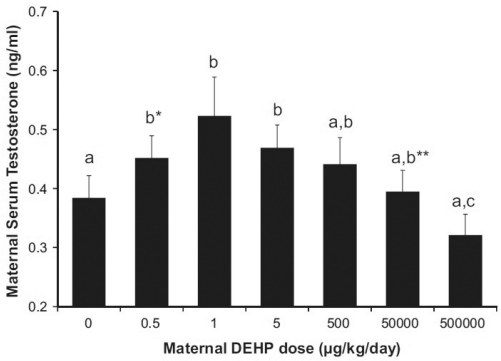


Fig. 2. Effect of different doses of DEHP on maternal serum testosterone concentrations on GD 18 (ANOVA on log-transformed data, $P < 0.05$). Values represent the mean ± SEM. Treatments with the same letter are not significantly different from each other but are statistically different from groups with other letters. b*, $P = 0.07$ relative to controls; b**, $P = 0.09$ relative to 500,000 group. Sample sizes were: oil $n = 20$; 0.5 μg, $n = 9$; 1 μg, $n = 11$; 5 μg, $n = 12$; 500 μg, $n = 13$; 50 mg, $n = 16$; 500 mg, $n = 17$.

FIGURE 1.2: Example of bar charts.

Although both representations allow a dense display, they have two major drawbacks: they assume normally distributed data (and we know how often this is violated in real data) and they do not allow access to the individual data. Individual datapoints have a special meaning in toxicology, because sometimes the relevant information is contained just in a few extreme values —not necessarily in means. Therefore, Rhodes et al. [319] visualized just group-specific individual data (by dots) for 20 rats together with the (geometric) mean (in Figure 1.3):

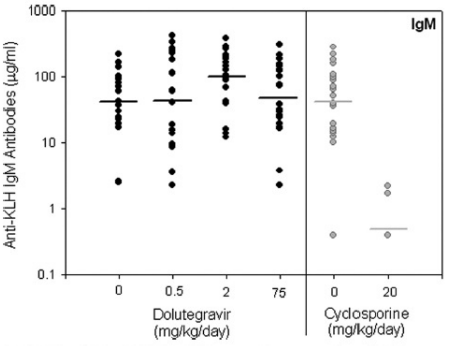


FIG. 2. IgM anti-KLH TDAR in juvenile rats treated with dolutegravir. Sera were harvested 5 days post-KLH immunization and assessed for anti-KLH IgM antibodies by quantitative electrochemiluminescent immunoassay. Circles represent individual animals (n = 20/group) with geometric mean indicated by the bars. There were no detectable dolutegravir-related effects on the anti-KLH IgM antibody response when juvenile animals received daily treatment. In rats given the positive control for immunosuppression (20 mg/kg/day of cyclosporine), there was a significant decrease ($p < 0.001$) in the level of anti-KLH IgM antibodies. Eighteen of 20 rats given 20 mg/kg/day of cyclosporine tested below the IgM assay lower limit of quantification and were assigned a value of 0.4 µg/ml for the purpose of calculating group geometric mean.

FIGURE 1.3: Example of individual data representation.

Surprisingly, boxplots are rarely used in toxicology (see e.g., [75] in Figure 1.4), although, they were recently recommended in a points-to-consider paper [108].

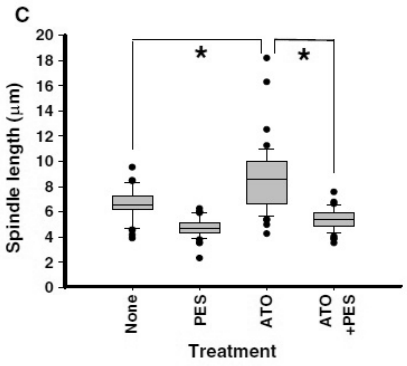


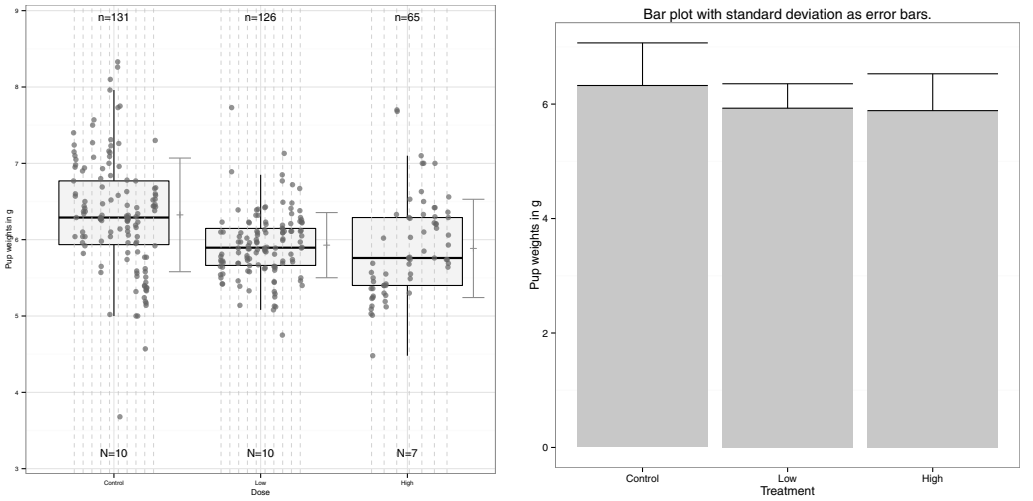
FIGURE 1.4: Example for boxplot in toxicology.

A boxplot typically uses nonparametric measures of location and scale, namely median and interquartile range ($IQR = q_{0.75} - q_{0.25}$) as well as an outlier rule (represented by whiskers), e.g., $k * IQR; k = 1.5$ (notice, $q_{0.75}$ is the 75% percentile). The boxplot provides simple information on group-specific location, variance, and asymmetry of distribution as well as existence of extreme values. Outlier identification or even elimination is rather critical in safety assessment (see Section 2.5) and this k-rule should not be used without specific care [357]. For grouped data a specific jittered boxplot was developed in the `R library(toxbox)` (see details [294]) which provides: i) nonparametric measures median, IQR, ii) parametric measures mean and standard deviation, iii) individual data, iv) graphic differentiation between randomized unit and technical replicates (e.g., pregnant rats and pups), v) sample sizes, and vi) signs of significances. The inappropriateness of barcharts compared to jittered boxplots (available in the package `toxbox`) is demonstrated by litter-specific pup weight data [390] in Figure 1.5 (see further details in Section 5.2).

```
> data("ratpup", package="WWGbook")
> Ratpup <-ratpup
> Ratpup$Treatment <-factor(Ratpup$treatment,
+   levels=c("Control","Low","High"))
> Ratpup$litter <- as.factor(Ratpup$litter)
```

This specific version of boxplots is used throughout this book for grouped data; here is the R-code:

```
> library("toxbox")
> boxclust(data=Ratpup, outcome="weight", treatment="Treatment",
+ cluster="litter", ylabel="Pup weights in g", xlabel="Dose",
+ option="dotplot", vline="fg", hjitter=0.01, legpos="none", printN=TRUE,
+ white=TRUE, titlesize=8, labelsize=6)
> data.raw <-ratpup
> data.raw$value <-data.raw$weight
> data.summary <- data.frame(
+   treatment=levels(data.raw$treatment),
+   mean=tapply(data.raw$value, data.raw$treatment, mean),
+   n=tapply(data.raw$value, data.raw$treatment, length),
+   sd=tapply(data.raw$value, data.raw$treatment, sd)
+ )
> data.summary$treat <-factor(data.summary$treatment,
+                       levels=c("Control","Low","High"))
> library("ggplot2")
> ggplot(data.summary, aes(x = treat, y = mean)) +
+   geom_bar(position = position_dodge(), stat="identity", fill="gray") +
+   geom_errorbar(aes(ymin=mean, ymax=mean+sd)) +
+   ylab("Pup weights in g") +
+   xlab("Treatment") +
+   ggtitle("Bar plot with standard deviation as error bars.") +
+   theme_bw() +
+   theme(panel.grid.major = element_blank())
```



(a) Boxplot (b) Bar chart

FIGURE 1.5: Comparing two graphical representations.

The bar chart visualizes just means and standard deviations (SD) (assuming normal distribution), whereas in the boxplots the raw data, parametric and nonparametric summary measures (namely means, standard deviation; median, interquartile range), the individual data (dots), and the per-litter structure (including group-specific number of litters and the number of plots) can be seen. For incidences or graded histopathological findings mosaic plots can be used for visualization; see e.g., Section 2.1.8.

1.2.2 How to present tests' outcomes: Stars, letters, *p*-values, or confidence intervals?

Four different presentation styles of statistical tests can be found in the literature: i) rejection/non-rejection of H_0 (where letters indicate non-distinguishable treatment groups; see an example below), ii) rejection of H_0 for three α levels (0.05, 0.01, 0.001) visualized by stars $*, **, ***$, iii) p-value (p), and iv) confidence interval. Stars are still common when results of chronic toxicity studies are presented; see e.g., [119] for selected hematological endpoints. A p-value is a near-to-zero skewed probability, reflecting the falsification principle: *we can never prove an effect directly, only demonstrate how unlikely its opposite is.* The p-value is a measure of this unlikeliness as a probability between $[0, 1]$ [100]. A confidence interval provides the possible range of a particular effect size (such as $\mu_i - \mu_C$ or μ_i/μ_C), for example, 95% of future data (where μ_i is the expected value in group i). The width of the confidence interval is a measure for the uncertainty, depending on variance(s), sample size(s), the false positive decision probability, and possibly other aspects of the assumed statistical model and test procedure. It contains information on: i) the rejection/non-rejection of H_0 by inclusion/non-inclusion of the value of H_0 (i.e., 0 for difference and 1 for ratio), ii) the interpretation of biological relevance by the distance of the confidence limit(s) to this value of H_0 in terms of the measured unit (difference) or percentage change (ratio), and iii) the directional decision (i.e., whether increasing or decreasing effects occur). Therefore, confidence intervals should be preferred [99, 26], particularly their consistent use for

both proof of safety and proof of hazard [28]. In the ICH E9 guideline [198] their use is recommended for the evaluation and interpretation of randomized clinical trials; a related recommendation for toxicological studies is still missing. Today the most common presentation style in this field is the p-value due to its simplicity. Due to its smallness, the distance to H_0 in terms of a probability characterizes the magnitude of statistical significance. Summarizing, this book focuses on test decisions primarily by means of confidence intervals, and secondarily by compatible tests and their p-values. Therefore, the primary criterion should be the selection of an adequate effect size, namely for the difference of expected values $\mu_i - \mu_0$, the ratio μ_i/μ_0, the difference of proportions $\pi_i - \pi_0$, the ratio of proportions (relative risk) π_i/π_0, the odds ratio of proportions $\frac{\pi_i/(1-\pi_i)}{\pi_0/(1-\pi_0)}$, and the relative effect size $p_{i0} = Pr(X_i < X_0) + \frac{1}{2}Pr(X_i = X_0)$ (where X_i is a value in group i) (see details in Section 2.1.4)—appropriate for toxicological reasoning for a particular endpoint.

The joint visualization of data characteristics and statistical significances is a challenge. Here, different styles are demonstrated using an example of triglycerides in clinical chemistry for a control and three doses in a 13-week study with sodium dichromate dihydrate to F344 rats [16]. The raw data are presented in Table 1.1.

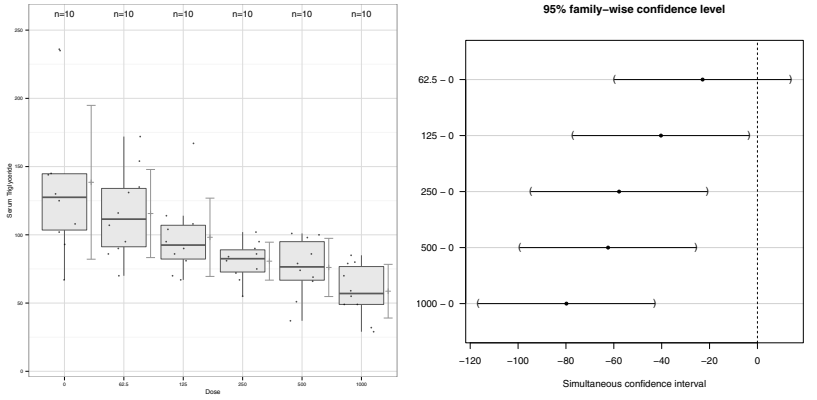
```
> data("clin", package="SiTuR")
```

Table 1.1: Raw Data of Serum Triglyceride of a 13-Week Study

Dose	Triglyceride
0	236.00
0	102.00
0	144.00
0	130.00
0	125.00
0	93.00
0	235.00
0	67.00
0	145.00
...	...
1000	85.00
1000	80.00
1000	29.00
1000	55.00
1000	59.00
1000	32.00
1000	79.00
1000	70.00

A possible visualization is a factorplot [32] which contains all-pair comparison adjusted p-values with the estimates and errors in a matrix. Because such a matrix is helpful when considering many all-pairs comparisons it is not demonstrated here.

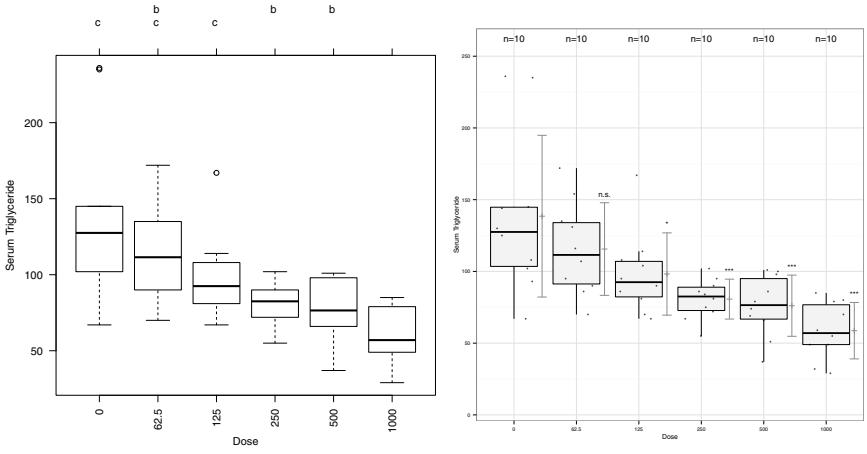
Multiple plots are preferred, with Figure 1.6a containing on the left side jittered boxplots and on the right side simultaneous confidence intervals for Dunnett-contrasts using the package `multcomp` (Figure 1.6b). A sometimes used visualization style is a letter plot for distinguishable treatment groups together with boxplots [304]. The package `multcompview` is a realization for all-pair comparisons whereas, in Figure 1.7a, the function `cld` (compact letter display) in the package `multcomp` is used for any multiple contrasts. Notice,



(a) Boxplots (b) Dunnett-type confidence interval

FIGURE 1.6: Multiple plots for serum triglyceride data.

throughout the book 95% confidence intervals are used and if not not will be stated clearly. Alternatively on the right side, a combination of boxplots with Dunnett-type p-values (Figure 1.7b) (or in the form of stars; see e.g., Section 1.2.2) [294] is displayed. This style allows for a maximum of interpretation.



(a) Letter-type representation of significances (b) Boxplots with Dunnett-type stars

FIGURE 1.7: Two plots with signs of significance for serum triglyceride data.

1.2.3 Proof of hazard or proof of safety?

The main objective of a study in regulatory toxicology is to decide whether a new drug entity or chemical is harmless, harmless up to a specified dose, or harmful. The final determination regarding the harmlessness or harmfulness is a complex process where much more is needed than just statistical tests.

1.2.3.1 Decision making in toxicology

First, the p-value of a statistical test is widely used in decision making: a finding is classified as positive when $p < 0.05$ or negative when $p > 0.05$. As an example, the criteria for positive response in the OECD 486 guideline for the UDS assay (unscheduled DNA synthesis) is defined as " *NNG (net nuclear grain) values significantly greater than concurrent control*" [289]. However, common statistical tests are formulated for a point-zero null hypothesis, such as the t-test, Wilcoxon–test, or χ^2-test, i.e., each tiny non-relevant change versus control can be significant, for example, in the case of small variances. Conversely, large effect sizes can be deemed "not significant" simply because the sample size was too small for the inherent variability of the particular endpoint.

Second, a modified decision rule is used that takes the magnitude of the spontaneous tumor rate into account. The two significance levels for a positive Cochran–Armitage [31] trend test for proportions are: $p < 0.025$ for rare tumors and $p < 0.005$ for common tumors [14]. This rule reflects the specific sensitivity of tests on proportions depending on the spontaneous proportion.
Third, trend tests are used for claiming a positive trend because the common designs in toxicological assays include a negative control and several doses. For example, in the OECD 479 guideline for the *in vitro* sister chromatid exchange assay (SCE) a criterion is formulated as: *"... statistically significant dose-related increase in the mean number of SCEs per cell"* [11].
Fourth, relevance thresholds are used, e.g., in the local lymph node assay (LLNA) a positive response is concluded when the stimulation index is larger than 3 [286]. Most relevance criteria are formulated for relative change, such as the k-fold rule in the Ames assay [70], and therefore inference relative to control is needed. However, these relevance thresholds ignore the assay's uncertainty. Here, the estimation of confidence intervals for ratio-to-control comparisons is proposed, where their estimated limits are used *posthoc* to interpret relevant (proof of hazard; see Sections 1.3, 2.1) or tolerable change (proof of safety; see Section 2.5).

1.2.3.2 Be confident in negative results

Due to the falsification principle of Neyman–Pearson tests, two concepts should be distinguished: i) the proof of hazard, i.e., demonstrating harmfulness and ii) the proof of safety, i.e., demonstrating harmlessness. In the first approach the probability of erroneously concluding hazard, i.e., producer's risk (false positive error rate), will be controlled directly. In the second approach the probability of erroneously concluding safety, i.e., consumer's risk (false negative error rate $f-$) is controlled directly [58, 262]. (Notice, alternatively the term power is used, whereas power is simply $1 - f-$.) For drug safety assessment, the consumer's risk seems to be more important, along the lines of: *"be confident in negative results"* [181, 260]. Failing to reject the null hypothesis in the proof of hazard ($p > 0.05$), often leads to the conclusion of evidence in favor of harmlessness. However, it is important to realize: *"absence of evidence is no evidence of absence"* [27, 28]; this can lead to problematic decisions. Therefore, both the widely used proof of hazard and the rarely used proof of safety are described in the following; see particularly Sections 2.1 and 2.5.

1.2.3.3 Several two-sample comparisons or multiple comparisons versus control?

For the common design, which includes a control and several doses, the US National Toxicology Program (NTP) [9] recommends analysis by the Dunnett procedure [102] and/or Williams procedure [400]. These proof of hazard approaches control the familywise error rate (FWER). The term *familywise* is a bit difficult to understand —*family* consists exactly of the k comparisons against control for k dose or treatment groups. A more appropriate term is *claimwise error rate* [301], where *claim* consists of the identification of at least one up to k doses/treatments different to control. Therefore, Dunnett/Williams procedure is rather conservative compared to independent level α-tests (see Section 1.3). Commonly, a significant dose-related trend test is used as causation criteria, e.g., in [290] *"...there are several criterion for determining a positive result, such as a concentration-related increase..."*. The commonly used test for comparing doses versus control under the assumption of monotonicity is the Williams procedure [400]. Therefore, this procedure is the main approach in this book and modifications used for the different endpoint types occurring in toxicology will be provided, e.g., for proportions, counts, poly-3 counts, survival functions, and ranks. Evaluating an efficacy endpoint in a clinical dose finding study by Dunnett's procedure is appropriate because the claim of a particular effective dose, a related α penalty should be paid. In contrast, within toxicology the aim is not to select a certain toxic dose and therefore the control of the comparisonwise error rate (CWER) seems to be sufficient. For this purpose, k independent two-sample comparisons $Cvs.D_i$, each at level α, are described in Section 1.3. This approach is used in order to achieve lower false negative rates (which is equivalent to higher power) compared to approaches controlling FWER. In order to show the difference between FWER and CWER, a fictional example containing the comparison of 3 doses vs. control is given based on the Bonferroni inequality (FWER control) in comparison to independent t-tests (CWER control) ($n_i = 20, \sigma/\delta = 1$, two-sided t-tests)($\sigma$ ... root of variance, δ ... difference to be detected). The main differences are the elementary α levels of $\alpha^{CWER} = 0.05$, and a much smaller $\alpha_i^{FWER} = 0.05/3 = 0.01666$. Both false negative decision rates can be simply calculated:

```
> fnC <-1-power.t.test(n =20, delta =1, sd = 1, sig.level = 0.05)$power
> fnF <-1-power.t.test(n =20, delta =1, sd = 1, sig.level = 0.01667)$power
```

Consequently, the *posthoc* false negative error rates for the CWER approach of 0.131 is much smaller than those of the FWER approach (0.258). This illustrates a basic dilemma in toxicology caused by applications of Dunnett-type evaluation in most reports and publications. (Please notice, that throughout this book the correlation between test statistics is used, therefore all FWER-approaches, such as Dunnett test, are less conservative than the Bonferroni approach which is only used here to keep the example simple.)
For a particular toxicological assay, statistical significance or biological relevance is rarely explicitly defined. The second part of the OECD recommendation [290] focuses on biological relevance, i.e., *"...a reproducible increase in the number of cells containing micronuclei"*. Therefore, related simultaneous confidence intervals will be described primarily. For some assays, adjusted p-values will be reported in order to achieve adaptability to common publication style (dominated by p-values). In the following, two types of proof of hazard are discussed: without and with control of FWER. In Section 1.3 independent two-sample comparisons will be discussed, whilst in Section 2.1 the Dunnett- and Williams-type procedures for multiple comparisons versus control will be discussed. The main focus here is the Williams procedure for several types of endpoints. Simultaneous confidence intervals, allowing a claim for relevance, are of primary interest.

1.2.4 Sample size matters

In most guidelines the minimum required sample size for a toxicological assay is specified, e.g., at least triplicate plates in the Ames assay [287] or at least 5 animals per sex in the micronuclei assay (MN) [288]. For point-zero hypotheses in the proof of hazard, the secondary error rate (false negative rate f^-) is therefore endpoint-specific, depending on the variance, the common $\alpha = 0.05$ level and several assumptions such as Gaussian distribution. Still, the minimum required sample sizes ensure a certain standardization of the false negative rate ($power = 1 - f^-$) and should be used in practice. In the proof of safety, the secondary error rate depends additionally on the endpoint-specific tolerance threshold θ_i (see Section 2.5). Another aspect regarding relatively small sample sizes, such as triplicated plates in the Ames assay, is the specific finite properties of the tests. For example, asymptotic nonparametric tests do not control level α approaches, whereas permutation tests are rather conservative. Furthermore, the impact of variance heterogeneity is serious in both tests because both assume variance homogeneity.

Therefore, the small sample size behavior of the tests and procedures is considered in this book, as a specific feature of inference; see e.g., [46].

Outside of regulatory toxicology, sample sizes are sometimes chosen arbitrarily, i.e., not statistically planned in advance. The limitation of interpreting p-values becomes obvious, since the p-value is a direct function of sample size. In order to illustrate this, we generate simulated datasets containing the same expected values (means), and the same variances, but (balanced) sample sizes 5, 10, and 15.

```
> library(SimComp)
> set.seed(17059101); muC=7.6; muT=9.3; sC=1.54; sT=1.72
> exp5C <- ermvnorm(n=5,mean=muC,sd=sC)
> exp5T <- ermvnorm(n=5,mean=muT,sd=sT)
> exp10C <- ermvnorm(n=10,mean=muC,sd=sC)
> exp10T <- ermvnorm(n=10,mean=muT,sd=sT)
> exp15C <- ermvnorm(n=15,mean=muC,sd=sC)
> exp15T <- ermvnorm(n=15,mean=muT,sd=sT)
> endpoint5 <-c(exp5C, exp5T)
> factor5 <- factor(rep(c("Control", "Treatment"),c(5,5)))
> dat5 <- data.frame(endpoint5,factor5)
> endpoint10 <-c(exp10C, exp10T)
> factor10 <- factor(rep(c("Control", "Treatment"),c(10,10)))
> dat10 <- data.frame(endpoint10,factor10)
> endpoint15 <-c(exp15C, exp15T)
> factor15 <- factor(rep(c("Control", "Treatment"),c(15,15)))
> dat15 <- data.frame(endpoint15,factor15)
> p5 <-round(t.test(endpoint5~factor5, data=dat5)$p.value,3)
> p10 <-round(t.test(endpoint10~factor10, data=dat10)$p.value,3)
> p15 <-round(t.test(endpoint15~factor15, data=dat15)$p.value,3)
```

The Welch-t-test p-value for the small-sample size study is 0.139, for medium-sample size 0.032 and for larger-sample sizes 0.008. This example should warn against over-interpretation of p-values in unplanned studies.

A p-value depends, of course, primarily on the effect size, but also directly on the sample size (as well as the variance and other items such as the validity of the distribution assumption). The p-value has a predictive value in toxicological decision making in planned studies (*a*

priori assuming a minimum effect size) or at studies using guideline-recommended n_i, but not for studies where n_i results from non-statistical reasons such as feasibility.

1.2.5 Multiplicity occurs

For Phase III clinical trials the adjustment against several sources of multiplicity (multiple endpoints, multi-regional trials, subgroups, repeated studies) is essential [42, 301]. Thus, the concept of a claimwise error rate was proposed [301] which is more realistic than FWER. In safety assessment multiplicity-adjustment should be used with care due to its conservativeness. The standard design in toxicological studies includes a negative control and some doses: $C, D_1, ..., D_k$, and therefore the US National Toxicology Program proposed Dunnett [102] and/or Williams [400] tests [9], both controlling FWER. Moreover, multiple endpoints, such as multiple tumor sites occur and the question arises whether and how multiplicity adjustment should be performed. A quite different approach is used for the proof of hazard or the proof of safety. Both union-intersection (i.e., *OR* links between multiple hypotheses) and intersection-union hypotheses (i.e., *AND* links between multiple hypotheses) will be discussed. Along with this, several methods for multiplicity adjustment are discussed in this book.

1.2.6 Several types of endpoints occur

A contradiction may exist in toxicology. Whereas most publications on statistics in toxicology focus on continuous endpoints, such as organ weights, the most relevant information is contained in counts and proportions, such as tumor rates or graded histopathological findings, for which only few statistical approaches exist. Three main types of endpoints can be distinguished: i) vital signs, e.g., number of offspring per live female in *Ceriodaphnia dubia* assay [249], ii) pathological findings, e.g., number of micronuclei in the MN assay [176], and iii) physiological parameters, e.g., serum bilirubin concentration [19]. Endpoints measured by vital signs are used in inhibition assays, which are particularly suitable for ratio-to-control tests on non-inferiority, since their continuous variable, or count, or proportion decreases from large values in the control (sometimes 100%) to small values in the dose groups, i.e., the reduction of offspring is a sign of toxicity. In the second group, the data in the control are often zero or near-to-zero, requiring specific tests. For the third type of endpoints, increases or decreases may be of toxicological interest.

Several types of endpoints occur, e.g., continuous normal and non-normal variables, proportions, mortality-adjusted tumor rates, graded findings, counts, time-to-event data; and therefore corresponding endpoint-specific tests are discussed in this book.

1.2.7 Directional decisions

Although a controversy on the appropriateness of one-sided tests exists (see, e.g., [245]), one-sided tests for the first two categories of endpoints ((i.e., inhibition/non-inferiority endpoints and near-to-zero endpoints)) are clearly suggested. For both endpoints the opposite direction can be toxicologically irrelevant. For example, in a carcinogenicity study only increasing tumor rates are of interest; possibly decreasing tumor rates are toxicologically irrelevant. Moreover, one-sided hypotheses are inherently needed for non-inferiority tests (see Section 2.5.1). Furthermore, one-sided tests or confidence limits should be preferred to

limit the false negative error rate and are plausible for dose-related trend tests. Even for the third category of endpoints, a directional decision after a significant two-sided test is needed (which is closed under intersection) because not only a significant change but also its direction is of interest. Hereby, the control of the directional error rate is of particular interest, that is the rate for a decision in the false direction [122]. Please notice, appropriate one-sided tests and confidence intervals for differences of proportions for small sample sizes between $n_i = 10$ (90-days study) and about $n_i = 50$ (long-term carcinogenicity study) are a challenge. Due to their conservativeness, exact approaches, such as the exact Fisher test or permutation Wilcoxon–test, should be avoided in toxicology (see further discussions [248, 259]). Thus, approximate one-sided confidence intervals are provided for comparing treatments versus control, trend evaluation, and claiming non-inferiority. Please notice, related confidence intervals for ratio-to-control or odds ratio are unstable for near to zero spontaneous rates and should be avoided. Therefore, in this book one-sided confidence intervals (and their compatible tests) are of particular interest.

1.2.8 Specific designs

The common completely randomized design is $[C, D_{low}, D_{med}, D_{high}]$. Rarely, more or less dose groups are included. Balanced designs are recommended, i.e., the same sample size in each group. However, due to the power optimality rule of the Dunnett procedure, special unbalanced designs were proposed, i.e., using $\sqrt{k}$ higher sample size in the control. Notice, this rule remains valid only under variance homogeneity [93]. Further design issues are dual controls and positive controls. Dual controls should be avoided, as they are not required by guidelines, make variance heterogeneity more serious, and increase total sample size. Positive controls are highly recommended, first to guarantee assay sensitivity (by demonstrating a substantial effect size versus negative control), and second to demonstrate non-inferiority of dose effects relative to the positive control [37]. Unfortunately, the use of a selected positive control with a certain concentration is not common in toxicology. If both sexes are used, as in rodent studies, they are commonly evaluated independently (which is conditionally a conservative approach). Particularly, for large animals such as dogs and primates, the sexes are analyzed together to compensate for the rather small sample sizes. In some studies more complex layouts occur, including blocks (e.g., cages), a secondary factor (e.g., sex), hierarchical sub-units (e.g., pup weight within a female), technical replicates (e.g., 50 cells per gel and animal in the Comet assay), repeated measures (e.g., body weight curves), and covariates (e.g., body weight in organ weight analysis). Commonly simplified solutions are used (such as relative organ weights instead of covariance analysis), or per condition analysis is preferred (such as separate per-sex analysis) or block effects are ignored. Additionally, related appropriate approaches are available, focusing on the primary comparison of dose groups versus control, e.g., for repeated measures.

Therefore, the specification of a design with doses and a zero-dose control is primarily discussed in the book, extended for secondary fixed factors, sub-units, technical replicates, blocks, repeated measures, and covariates; see Section 2.4.

1.2.9 Mixing distribution and outliers

Most statistical textbooks assume unimodal distributions, though not necessarily normal. Nevertheless, the tolerance model of toxicology, in which *the affected population consists of τ responders and $(1 - \tau)$ non-responders* would lead to a mixing distribution, that is

not a unimodal distribution. However, a bimodal (or even a k-modal) distribution may exist, consisting of an unknown proportion $(1 - \tau)$ of animals behaving similarly to control animals and τ responding animals. Inference for mixing distributions exists, e.g., [170], but the common small sample size makes these approaches complicated (see Section 2.1.10). However a large number of sub-units occur in toxicology, such as scored cells or comet length for many cells in gel electrophoresis. Here, specific mixing distribution approaches can be used; see Section 4.12. Particularly, for small sample sizes it can happen that just one reacting animal cannot be distinguished from an outlier. Therefore, formal statistical tests on outliers should generally be avoided in toxicology (particularly in connection with their removal); a further reason is that the underlying distribution is unknown.

1.2.10 The phenomenon of conflicting decisions

In each animal several endpoints are measured or surveyed in repeated dose toxicity studies. Some are continuous, such as body weight, some are proportions, such as histopathological incidences, and some are ordered categorical, such as graded histopathological findings. For all these endpoints the decision is based on the same maximum false positive rate $\alpha = 0.05$ and the same sample sizes n_i. The detectable effects for an equal assumed false negative rate of say $\beta = 0.20$ are therefore extremely different. For example, we assume the body weight in $n_i = 10$ control rats is about 420 g and the standard deviation SD=14 g. For the one-sided t-test the detectable effect difference to a dose group is about 16 g. For a proportion with a spontaneous response rate of 0%, only an incidence difference to 45% in any dose group can be detected. For graded findings with 90% in category 0, 5% in category +, 2.5% in category ++ and +++ in the control, a shift to 40% in category 0, 20% in category +, 25% in category ++ and 15% in category +++ can only be detected [224]. This means that the same detection sensitivity is found: i) for an increase of 3.3% body weight, ii) for an increase of incidence from 0 to 45%, and iii) for a shift from 90% to 40% in category 0. This example characterizes the basic contradiction: small differences of precise measured continuous endpoints, which may have a low toxicological importance, can be detected. In contrast, only well-pronounced differences for either incidence rates or graded findings, which may have a high toxicological importance, can be detected.

Therefore, in this book appropriate and robust procedures for normal, continuous, proportions endpoints and counts are discussed and, in particular, their small sample behavior is discussed.

1.2.11 Decision tree approaches

One of the myths in applied biostatistics is that in the case of heterogeneous variances, the Wilcoxon–test (for two samples) or the Steel-type test (for comparing treatments vs. control, i.e., a nonparametric version of the Dunnett test) are robust, whereas, the t-test (or the Dunnett test) is not robust. For example in the method section of [407] it is explicitly written: *"...when the variance was heterogeneous based on Bartlett's test, the Steel's multiple comparison test was employed."* Therefore, decision tree approaches, consisting of a pre-test on variance homogeneity (to be precise: on variance heterogeneity) or other conditions, such as normal distribution, followed by the optimal selected main test, are quite common. Three major arguments exist against this conditional two-step approach, the so-called decision tree approach:

i) the conditional two-step approach does not control level α [419]: *"...preliminary tests of equality of variances used before a test of location are no longer widely recommended by statisticians, although they persist in some textbooks and software packages. The present study extends the findings of previous studies and provides further reasons for discontinuing the use of preliminary tests ",*

ii) nonparametric tests or procedures are inappropriate for heterogeneous variances [418]: *"...the Student t test maintains its significance level more consistently than the Wilcoxon–Mann–Whitney test when variances of treatment groups are unequal and sample sizes are equal,"*
iii) conditional tests are problematic [154]: *"...many books on statistical methods advocate a conditional decision rule when comparing two independent group means. This rule states that the decision as to whether to use a pooled variance test that assumes equality of variance or a separate variance Welch t test that does not should be based on the outcome of a variance equality test. Several unconditional tests including the separate variance test performed as well as or better than the conditional decision rule... conditional decision rule should be abandoned."*

Moreover, even for normal distributed variables, for unbalanced designs and variance heterogeneity, the Steel-type procedure tends to be conservative when high variances occur at large sample size and to liberal behavior (i.e., FWER is not controlled when high variances occur at small sample size). In the simulation study [271], for a design using $n_i = 35, 25, 20, 15$ and $SD_i = 1, 3, 3, 3$ the empirical FWER is 0.098 for the Steel procedure whereas for the parametric counterpart (Dunnett) it is 0.060; i.e., the decision tree assumption is seriously violated. Ruxton (2006) [329] recommended the Welch-t-test as an alternative to both the t-test and the Mann–Whitney U test. A further decision-tree problem is the use of a global ANOVA conditional or unconditional before the Dunnett procedure (or in the nonparametric setup the Kruskal–Wallis global test before the Dunn or Steel procedure), as described in the US NTP documents [9]. This is inappropriate for several reasons: i) from the view of the closed testing procedure after a significant F-test, level α t-tests "control vs. doses" form a closed intersection hypotheses system; it is not the procedure itself controlling the FWER. Performing the Dunnett procedure only after a significant F-test is an unnecessarily conservative approach, which should be avoided in toxicology [183], ii) the alternative regions of the F-test and the Dunnett procedure are different; the first is defined for an all-pairs alternative, the second for a simple tree alternative; and iii) the F-test is a quadratic form whereas the Dunnett test is a linear form as a special case of multiple contrast tests; see Section 2.1.1.1.2. Therefore, the F-test should not be used in a one-way layout at all. Rarely two- or higher-way layouts are used in toxicology (e.g., sex or time as a secondary factor), and here the F-test is used for testing interactions *a priori*. But even here, interaction contrast is a more appropriate technique [348]; see Section 2.4.1.

Therefore, in this book decision tree approaches are not proposed.

1.2.12 The special importance of control groups

Assays without a negative control group are unthinkable in toxicology. Therefore inference versus the concurrent negative control is the dominating approach; see Sections 1.3 and 2.1. Multiple similar assays allow the use of the information of their historical controls, either directly into the test decision (in Section 4.10) or the estimation of reference values (in Section 2.3). Positive controls are less common, but can be used either for claiming assay

sensitivity or for demonstrating the relevance of a change versus negative control by using a non-inferiority test (in Section 4.11).

Therefore, procedures for comparisons against the negative control ($C-$), procedures for comparisons against the positive control ($C+$), and the use of historical control group information will be discussed in this book.

The preferable design includes both a negative control, some dose groups and a positive control. It allows simultaneous decisions on superiority against $C-$ and non-inferiority against $C+$ (see details in Section 4.11).

As an example, micronucleus data with the treatment groups `Vehicle, Hydro30, Hydro50, Hydro75, Hydro100,Cyclo25` were selected [152] (see details in Section 4.11).

```
> data("Mutagenicity", package="mratios")
> Muta <- Mutagenicity
> Muta$Treatment <- factor(Muta$Treatment, levels=c("Vehicle","Hydro30",
+                            "Hydro50", "Hydro75","Hydro100","Cyclo25"))

> library("mratios")
> MutaP <-droplevels(Muta[Muta$Treatment !="Vehicle", ])
> pposs <-simtest.ratioVH(MN~Treatment, data=MutaP,
+                       alternative="less",  type = "Dunnett", base = 5,
+                       Margin.vec=c(0.8,0.8,0.8,0.8))$p.value.adj
> MutaN <-droplevels(Muta[Muta$Treatment !="Cyclo25", ])
> pnegs <-simtest.ratioVH(MN~Treatment, data=MutaN,
+        alternative="greater",  type = "Dunnett", base = 1)$p.value.adj
> MutaPN <-droplevels(Muta[Muta$Treatment %in% c("Vehicle", "Cyclo25"), ])
> ppn <-t.test.ratio(MN~Treatment, data=MutaPN, alternative="less")$p.value
```

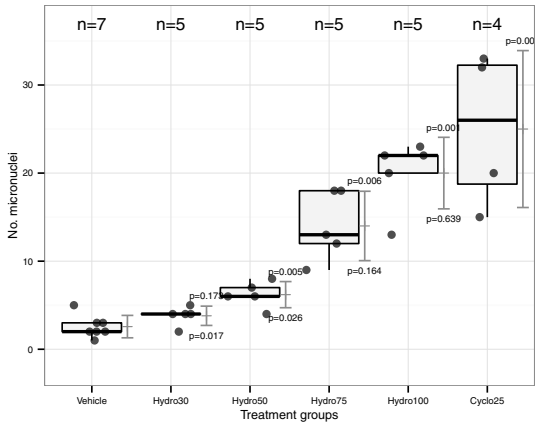


FIGURE 1.8: Micronucleus data including a positive control.

Figure 1.8 shows the jittered boxplots with adjusted p-values for Dunnett-type ratio-to-control comparisons: the lower for comparisons against $C-$, the upper for comparisons against $C+$. MNs in the Hydro100 dose are substantially increased against $C-$ (p-value

0.001) and 80% non-inferior to $C+$ (p-value 0.639). The p-value for the positive control stems from claiming assay sensitivity by a t-test for the two-sample problem $[C-, C+]$.

In some bioassays dual controls are used, commonly water and solvent control. The primary aim seems to be the identification of a possible specific effect of the solvent. This can be easily tested by two-sample tests described in Section 1.3. In many cases these controls do not really differ [132] and the possibility of pooling exists to increase control sample size and hence the power of the tests against control [131]. A pre-test is recommended to test for no serious difference between both controls. An equivalence test can be used; see Section 2.5. But the choice of the equivalence thresholds remains an open problem.

1.2.13 Statistical significance and biological relevance

A not uncommon practice in toxicology is to interpret selected significant findings as biologically not relevant. A less common, but more important, practice is to interpret insignificant findings as relevant. This contradiction between significance and relevance has several reasons, among them the use of tests for a point-zero null hypothesis. For example, for the UDS assay in the OECD 486 guideline *"...NNG values significantly greater than concurrent control"* are defined as positive outcomes. Confidence intervals should be used ([99, 279]) and interpreted with respect to an *a priori* defined relevance threshold, denoted as the minimal clinically important difference [208], e.g., 100 ml for lung function through FEV_1 (an expiratory volume parameter). Using a toy example, five outcome types can be distinguished, as can be seen in Figure 1.9 for a simulated dataset:

- Statistically not significant `D1-NC`
- Significant without clinical relevance `D2-NC`
- Not significantly less than threshold `D3-NC`
- Probably clinically significant effect `D4-NC`
- Large clinically significant effect `D5-NC`

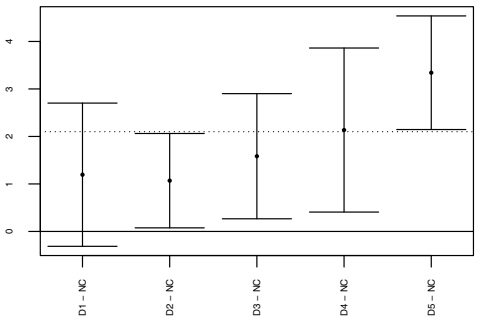


FIGURE 1.9: Simulated data for different patterns of significance and relevance.

From a toxicological point of view, the scenario *probably clinically significant effect* (`D4-NC`) is of particular interest, i.e., the effect size (here mean difference) is above the relevance threshold (i.e., certain biological relevance) and the lower confidence limit larger than zero

(i.e., formal statistical significant). For this scenario the sample size can be estimated [212]. For example, for an assumed effect difference $\delta = 3.0$, a standard deviation of $\sqrt{\sigma^2} = 2.3$, a false positive rate $\alpha = 0.05$, and a false negative rate $\beta = 0.2$, a sample size for the point-zero hypothesis (i.e., scenario of just statistical significance `D2-NC`) of $n_i = 8$ is needed (two-sided t-test with balanced sample sizes). For a scenario of significance and relevance (i.e., probably clinically significant effect (`D4-NC`)), a sample size of $n_i = 10$ is needed where the point estimator must be at least 2.1 (dotted line in Figure 1.9). These sample size estimations can be performed by the package `WinProb` using the concept of win probabilities [155].

A further argument for relevance is not only to consider a significant dose-response relationship for a particular endpoint in a single study, but also to consider a similar effect for multiple endpoints. Examples are negligible bias due to maternal toxicity in developmental studies, different dose-response patterns for different endpoints (e.g., malformations at low doses but fetal death at higher doses), taking into account cluster effects (i.e., a single malformation in 10 litters is not the same as 10 malformed pups in a single litter), similar effects for multiple studies and different species and supportive arguments from quite different mechanisms (e.g., toxicokinetics); see the rare definition of levels of evidence in developmental toxicology [117].

1.3 Proof of hazard using two-sample comparisons

1.3.1 Normal distributed continuous endpoints

In principle, the common two-sample tests, such as the t- or Wilcoxon–test, each at level α can be used here. As an example selected clinical chemistry data from the 13-week study with sodium dichromate dihydrate to F344 rats [16] is used; see part of the raw data in Table 1.2. In a `data.frame` format for a factor `Dose` and several endpoints (blood urea nitrogen (BUN), serum creatinin (Creat), albumin (ALB), serum glucose (SerumGlucose), creatine kinase (CreatKinase), anilin aminotransferase (ALT)) for each animal (i.e., row) complete values exist.

```
> data("clin", package="SiTuR")
```

```
> library("pairwiseCI")
> welcht <-pairwiseCI(value ~ dose, data=mclin, by="variable",
+         method="Param.diff", var.equal=FALSE, control="0")
> plot(welcht, lines=c(1,0.5, 2),  lty=c(1,2,2), CIcex=0.5,
+      cex.axis=0.28, cex.main=0.15, main=NULL)
```

Confidence intervals for Welch-t-tests can be used assuming normal distribution and allowing heterogeneous variances. However, they are scale-specific which makes a comparative interpretation complicated; see Figure 1.10. Therefore, this subsection focuses on confidence intervals for ratio-to-control comparisons in order to allow claiming biological relevance (vs. statistical significance) over differently scaled multiple endpoints. Here the effect size is percentage change which is dimensionless. Two-sample one-sided or two-sided confidence intervals for the ratio to control according to Fieller [116] are available modified in the case of variance heterogeneity [367]. These intervals allow scale-independent interpretation, which

Table 1.2: Clinical Chemistry Raw Data of Sodium Dichromate Bioassay

Dose	BUN	TP	ALB	SerumGlucose	CreatKinase	ALT
0	17.3	7.4	5.0	137.0	202.0	100.0
0	15.0	6.9	4.7	180.0	205.0	56.0
0	15.7	7.1	4.9	164.0	188.0	70.0
0	16.5	7.1	4.9	145.0	155.0	54.0
0	16.7	7.2	5.0	126.0	160.0	64.0
...	...	...	...	...	...	...
1000	17.4	5.7	4.0	142.0	444.0	182.0
1000	17.4	6.2	4.3	125.0	401.0	118.0
1000	18.7	6.1	4.2	141.0	337.0	222.0
1000	20.5	6.1	4.2	146.0	838.0	322.0
1000	19.0	6.7	4.7	113.0	370.0	288.0

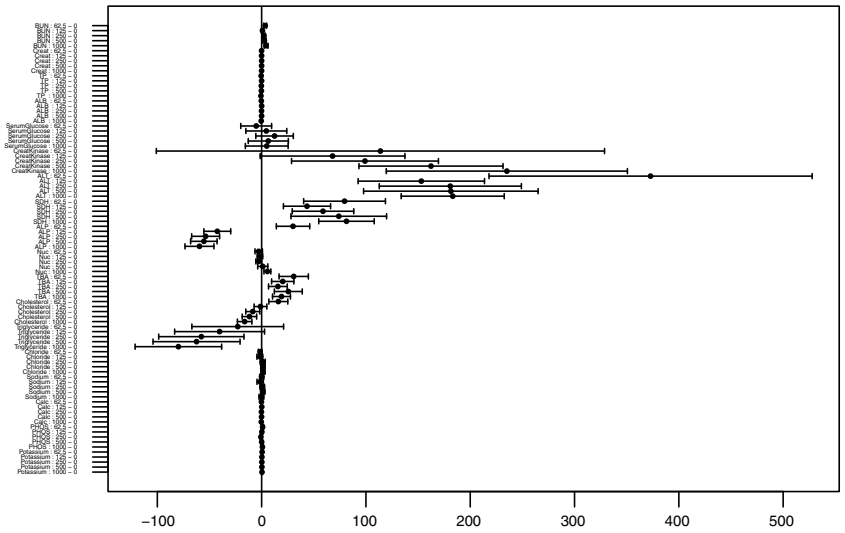


FIGURE 1.10: Welch-t-test confidence limits for clinical chemistry endpoints.

is particularly helpful for multiple endpoints in hematology and clinical chemistry. Naturally, this approach is limited to data containing control values different from zero, where variances and sample sizes are taken into account. By means of the R package `pairwiseCI` for all clinical chemistry endpoints, related dose-specific confidence intervals can be easily estimated. The numeric variable `Dose` should be transformed into the factor `Dose` and the row-wise structured multiple endpoints should be transformed into a secondary factor `variable` by the function `melt`. These two-sided confidence intervals for ratio-to-control comparisons allowing variance heterogeneity, independent of each endpoint and each dose, are presented in Figure 1.11. The related R-code is simple:

```
> library("pairwiseCI")
> tratio <-pairwiseCI(value ~ dose, data=mclin, by="variable",
+         method="Param.ratio", var.equal=FALSE, control="0")
> plot(tratio, H0line=c(1,0.5, 2),  H0lty=c(1,2,2), CIcex=0.5,
cex.axis=0.29, main=NULL)
```

We see that confidence limits of most endpoints and most comparisons are within a range between $[0.5; 2]$, but some endpoints, such as ALT, reveal a strong and statistically signif-

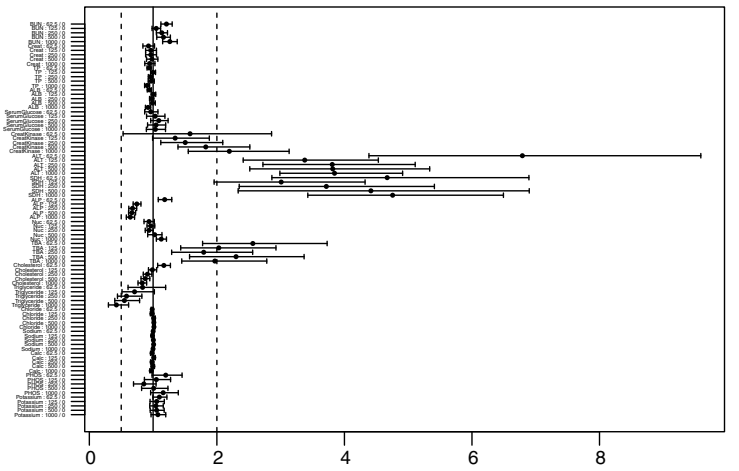


FIGURE 1.11: Ratio-to-control confidence intervals for clinical chemistry endpoints.

icant increase. The task of the toxicologist is to interpret these significant increases and to determine if they are toxicologically relevant,particularly in the context of all available findings. Please notice, this plot is suitable for screening non-normal findings because of its rather liberal behavior due to only comparison-wise error control between multiple doses and between multiple endpoints.

Alternatively, related p-values or confidence intervals for the differences to control can be estimated easily using the R package `pairwiseCI` using the option `method="Param.diff"` (see Figure 1.10).

1.3.2 Log-normal distributed continuous endpoints

A usual approach for the often skewed biological data is to apply a log-transformation. The back-transformed confidence intervals of the t-test represent as an effect size the ratio of medians which are hard to interpret. A more appropriate approach for log-normal distributed endpoints is available [74]. Even more extreme conditional tests based on data transformation are used in toxicology as a third alternative, i.e., [85] *"...if Bartlett's test was not significant at the* 1% *level, then parametric analysis was applied. If Bartlett's test was significant at the* 1% *level then logarithmic and square-root transformations were performed and if Bartlett's test was still significant, then nonparametric tests were applied."* Such an approach cannot be recommended; see the arguments in the decision tree Section (1.2.11).

Related confidence intervals can be estimated using the parameter `methods="Lognorm .ratio"` in the package `pairwiseCI` [338]:

```
> library("pairwiseCI")
> par(mar=c(4,8,3,1))
> lognorm <-pairwiseCI(value ~ dose, data=mclin, by="variable",
+         method="Lognorm.ratio", var.equal=FALSE, control="0")
> plot(lognorm, H0line=c(1,0.5, 2),  H0lty=c(1,2,2), CIcex=0.5,
+ xlab="Ratios-to-control for log-normal endpoints",cex.axis=0.29,main=NULL)
```

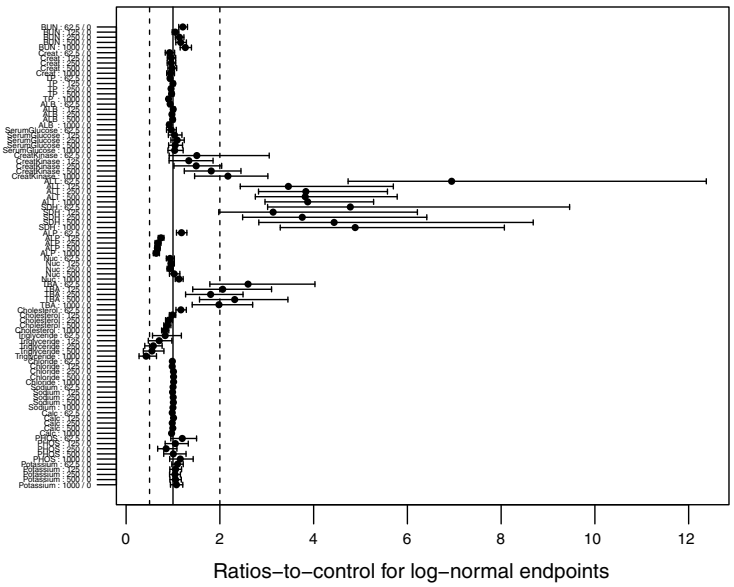


FIGURE 1.12: Confidence limits for log-normal distribution.

1.3.3 Non-normal distributed continuous endpoints

The common approach for endpoints with a skewed distribution and/or extreme values is the Wilcoxon–test. Sometimes a Behrens–Fisher problem occurs, i.e., the two groups differ primarily according to their locations but heterogeneous variances occur as well. Please notice, the Wilcoxon rank sum test is not suitable for variance heterogeneity, particularly in unbalanced designs [210]. There is a common related misuse; see e.g., evaluation of the chronic toxicity study on 1,3-dichloropropene [362]. Alternatives, such as the asymptotic rank-transformed Welch-t-test [420] cannot be recommended because simulations reveal inflated α levels [89] which are similar to their bootstrap modification [317] and to their α inflation under heteroscedasticity [378]. Despite its common use, data transformation to achieve variance homogeneity, (e.g., log transformation), is dangerous because transformation of the data prior to testing necessarily changes the null hypothesis under test [402]. Furthermore, the median test cannot be recommended because of its low power [65]. Adaptive tests consisting of several tests for several alternatives [278] are not recommended here because their alternative is not clearly defined and hence its toxicological interpretation is problematic and confidence intervals are not available. A particular alternative is the asymptotic nonparametric Behrens–Fisher test [62]. However, this test does not control the α level for small sample sizes and therefore modifications were introduced: a t-distributed version [62] and a permutation version [277]. By using the relative effect size [62], the null hypothesis $H_0 : p = \frac{1}{2}$ can be tested for two independent random variables X_1 and X_2 with any distributions F_1 and F_2 to: $p_{12} = Pr(X_1 < X_2) + \frac{1}{2}Pr(X_1 = X_2) = \int F_1 dF_2$. That is, p_{12} represents the probability that a randomly chosen subject in treatment group 1 reveals a smaller response value X_1 than a randomly chosen subject from treatment group 2 with response value X_2. If $p < 1/2$, then the values in group 1 tend to be larger than those in group 2. If $p = 1/2$, none of the observations tend to be smaller or larger. In the special case of independent ordinal data, p_{12} is called the ordinal effect size [331, 330]. Parametric tests use as effect size the difference (or ratio) of means, but for biomedical trials an effect size on an individual basis is of interest [61], such as the relative effect size. In summary,

all these arguments for and against nonparametric tests in comparison to parametric tests make the decision for one or the other complicated, this is true especially in small sample designs [29]. Nonparametric intervals for the difference to ratio according to [163] and for ratio to control [271] can be used. The related confidence intervals can be estimated using the parameter `methods="HL.ratio"` (where HL stands for Hogdes–Lehmann [163], specific nonparametric estimates):

```
> library("pairwiseCI")
> mclin0 <-droplevels(mclin[mclin$variable!="Creat", ])
> HLL <-pairwiseCI(value ~ dose, data=mclin0, by="variable",
+                  method="HL.ratio", control="0")
> mclin0 <-droplevels(mclin[mclin$variable!="Creat", ])
> par(mar=c(4,8,3,1))
> plot(HLL, H0line=c(1,0.5, 2),  H0lty=c(1,2,2),  CIcex=0.3,
+ xlab="Nonparametric ratio-to-control comparisons",
+ cex.axis=0.22, main=NULL)
```

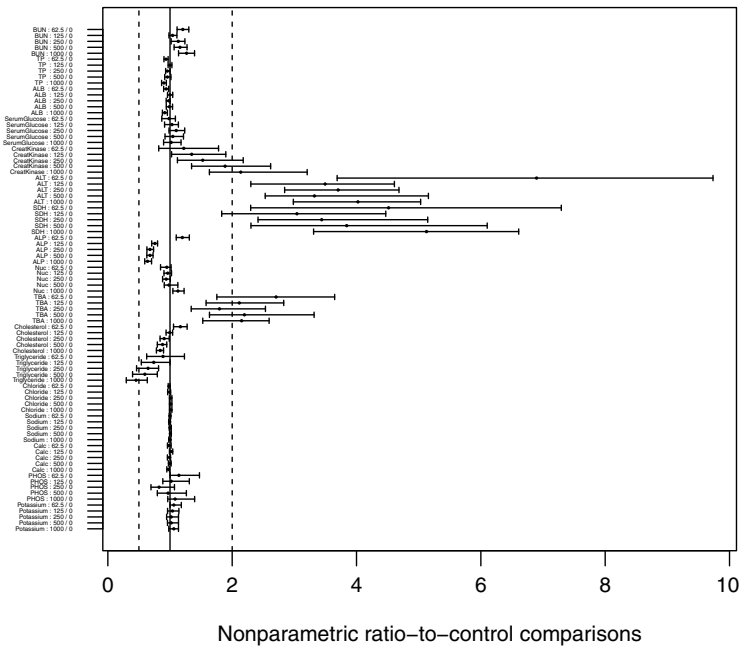


FIGURE 1.13: Nonparametric confidence limits for clinical chemistry endpoints.

1.3.4 Proportions

Up to now, an almost confusing number of publications on confidence intervals (and tests) for proportions in 2-by-2 tables (e.g., [257]) exists. Even with these, tests or confidence intervals for proportions in both 2-sample and k-sample designs are still a challenge. In toxicology for the analysis of crude mortality, findings or animal specific MN/PCE proportions (MN/PCE ...number of micronuclei per scored polychromatic erythrocytes), etc. rather specific conditions occur:

- Extremely small samples sizes (see 1.2.10). (Notice, $n_i = 10$ is *small* for the asymptotic Wilcoxon–test, but *extremely small* for a test on proportions)
- Both two-sided and one-sided hypotheses are of interest. However, switching from two-sided to one-sided intervals is not as simple as switching the quantile in t-test intervals $t_{df,1-\alpha/2} \Rightarrow \ t_{df,1-\alpha}$. One-sided limits may not provide a valid level $\alpha/2$ decision in situations where the two-sided intervals provide a level α decision. Even the approaches for lower limits, needed in the proof of hazard, and for upper limits, needed in the proof of safety, are not always the same [340].
- Zero or near-to-zero proportions in the control may happen, particularly for pathological responses such as tumor incidences. This data condition may cause serious problems when using a GLM-approach (see for details Section 2.1.5.1)
- Three different effect sizes for the expected proportions $\pi_i = Y_i/n_i$ are available and are used, though are not necessarily appropriate (where Y_i is the absolute number of effects). Although the risk difference $(\pi_i - \pi_0)$ is dominating in toxicology, the risk ratio (π_i/π_0) and the odds ratio $(\frac{\pi_i/(1-\pi_i)}{\pi_0/(1-\pi_0)})$ can be used as well. Actually, the choice of effect measure should depend on the design and interpretability, but it depends also on the data conditions and the numerical stability of the estimates. Risk ratio approaches are unstable when either the control rate is zero or the treatment rate is near to one and therefore should be used with care.
- Proportions occur on the level of the treatment factor, e.g., number of tumors in the high dose (with respect to number at risk), or on the level of each individual animal, e.g., animal-specific MN/PCE. In the first case, the 2-by-2, or 2-by-k table data can be analyzed, in the second case the variability between the animals should be modeled: either by estimation of the overdispersion (see Section 2.1.5.4) or by a mixed effect model with the random factor animal (see Section 2.1.5.4.1)
- Modified proportions occur, e.g., the poly-3 estimates by mortality-adjustment of tumor incidences (see 3.5)
- Stratified tables exist, e.g., for the analysis of incidental tumors (see Section 3.4.1)
- Tests and confidence intervals are not necessarily compatible

Although widely used, Fisher's exact test can be too conservative for small sample sizes [83], a contraindication for toxicology. On the other hand, approximate Wald-type intervals may seriously violate the coverage probability of 95%, particularly in case of near-to-zero control data. Therefore, the use of approximations can be suggested where *adding one pseudo-observation to success and to failure* [22] (denoted as ADD-2 approach), i.e., the Wald-type proportion $p_i = Y_i/n_i$ is replaced by $\tilde{p_i} = (Y_i + 1)/(n_i + 2)$. This is simple and efficient for risk differences [346] and for odds ratios [234] to control the coverage probability approximately. Please notice, for one-sided lower limits even adding one pseudo-observation, i.e., $\tilde{p_i}^{one-sided} = (Y_i + 0.5)/(n_i + 1)$ is appropriate [340, 346]. These interval estimates are numerically available in the R package `pairwiseCI` [341]. As an example histopathological tubular epithelia findings in the P-Cresidine carcinogenicity study [201] are used here, where 2-by-2 table data were selected (see Table 1.3).

Table 1.3: 2-by-2 Table Data for Tubular Epithelia Findings

Groups	Without	With
Control	10	0
High	2	8

```
> data("tubepi", package="SiTuR")
> tub2 <-droplevels(tubepi[tubepi$Group %in% c("Control", "High"),])

> library("pairwiseCI")
> tubOR <- pairwiseCI(cbind(TubularEpithelia,Without) ~ Group, data=tub2,
+                     alternative="greater", method="Prop.or",
+                     CImethod="Woolf", control="Control")
> tubD <-pairwiseCI(cbind(TubularEpithelia,Without) ~ Group, data=tub2,
+                     alternative="greater", method="Prop.diff",
+                     CImethod="AC", control="Control")
> tubRR <-pairwiseCI(cbind(TubularEpithelia,Without) ~ Group, data=tub2,
+                     alternative="greater", method="Prop.ratio",
+                     CImethod="MOVER", control="Control")
```

The difference of proportion is 0.8 where a significant increase with a magnitude of 0.423 in terms of the difference of proportions (95% lower limit) can be stated. The ratio of proportions cannot be estimated. The odds ratio is 71.4 where a significant increase with a magnitude of 5 in terms of an odds ratio was found.

1.3.5 Counts

Counts do not occur too frequently in systemic toxicology, as graded histopathological findings are common. In contrast, counts are the typical endpoints of *in vivo* or *in vitro* assays, such as number of micronuclei (MN) (see Section 4.6). Counts are also very common in sublethal ecotoxicity data, such as number of *Lemna fronds* or number of offspring for invertebrates. Already mentioned, one-sided testing is appropriate, since only an increase of severity is of interest. Three serious problems occur when comparing two counts of small sample sizes, such as ($n_i = 10, 10$): i) the power is rather small (compared with continuous endpoints), ii) the asymptotic tests for comparing two counts do not control the α level, and iii) the count measured per animal varies within each treatment group, i.e., an excess variability may exist. Related confidence intervals for the ratio of two counts exist by: i) fitting a generalized linear model with family definition of `quasi-poisson` using the function `glm` by constructing a deviance profile and deriving an equal-tailed confidence interval from this profile, ii) fitting a generalized linear model with log-link using the function `glm.nb` in package `MASS` by constructing a likelihood profile and deriving an equal-tailed confidence interval for the negative binomial model. Both intervals are available in the package `pairwiseCI` [341].

As an example, the two-sample problem with both a negative and a positive control for the number of micronuclei (`Vehicle,Cyclo25`) is used (see details in Sections 4.11 and 1.2.12). The related confidence interval can be estimated assuming Poisson-distributed counts with overdispersion between the animals using the parameter `methods="Quasipoisson.ratio"`:

```
> library("pairwiseCI")
```

```
> hyaCI <-pairwiseCI(MN ~ Treatment, data=MutaPN, alternative="greater",
+                    control="Vehicle", method="Quasipoisson.ratio")
> hyaCI3 <-as.data.frame(hyaCI)
```

For this ratio-to-control comparison, the ratio is 9.72 and its lower confidence limit is 5.99, i.e., the number of micronuclei is more than doubled, i.e., this assay is sensitive. Please notice the limited precision of this confidence limit because of the rather small sample sizes.

1.3.6 Further endpoint types

The analysis of multinomial variables (such as differential blood count) (see Section 2.1.7), ordered categorical data (such as graded histopathological findings) (see Section 2.1.8), hazard rates (such as mortality functions) (see Section 3.2.2), transformed endpoints (such a transformed near-to-zero counts) (see Section 4.7), censored time-to-event endpoints (see Section 3.6.4) and multivariate endpoints (see Section 3.6.2) are described in the related chapters in detail.

Simultaneous comparisons of dose or treatment groups against a negative control group is THE dominating approach for both short-term and long-term bioassays. This approach is described in detail in a separate chapter.

References

[1] Lung alveolar cell adenoma in male mice tumor data provided by Dr. Atiar Rahman of Division of Biometrics 6, Office of Biostatistics, CDER, FDA. 2007. Technical report.

[2] National Toxicology Program. 13 Weeks gavage study on female B6C3F1 mice administered with acrylonitrile (107-13-1, C50215B,5021501). Organ weights. Technical report.

[3] National Toxicology Program. 13 Weeks gavage study on female F344 rats administered with sodium dichromate dihydrate (VI) (CASRN: 7789-12-0, Study number: C20114,TDMS number:2011402. Technical report.

[4] National Toxicology Program. 2 Years bioassay of mercuric chloride on female rats. Body weight data. 1993 (study number C60173). Technical report.

[5] National Toxicology Program. 28 Days immunotoxicity bioassay on mice treated with Chloramine (2000). Technical report.

[6] National Toxicology Program. General toxicology. Short term (05161-03) toxicity evaluation of riddelliine (23246-96-0) on F 344/N rat. Technical report.

[7] National Toxicology Program. Micronucleus assay (A63788) of 5-(4-nNitrophenyl)-2,4-pentadien-1-al (NPPD) - 2608482. Technical report.

[8] National Toxicology Program. Reproductive toxicology. Diethylene glycol dimethyl ether. Study number: TER85061 on Swiss CD-1 Mice. Technical report.

[9] National Toxicology Program. Statistical procedures. Expanded overview (2013). Technical report.

[10] National Toxicology Program. Toxicology and carcinogenesis studies of mercuric chloride (CAS No. 7487-94-7) in F344 rats and B6C3F1 mice. No.408. Technical report.

[11] Test No. 479: Genetic toxicology: In vitro sister chromatid exchange assay in mammalian cells. OECD 23 Oct 1986. Technical report.

[12] U.S. Environmental Protection Agency Office of Water (4303T). Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms, Fourth Edition. Section 13: Test method daphnid. Survival and reproduction method. Table 4. Technical report.

[13] National Toxicology Program. Bioassay of piperonyl butoxide for possible carcinogenicity (CAS No. 51-06-6 / NCI-CG-TR-120). Technical report, 1979.

[14] Guidance for industry: Statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals. Technical report, US Food and Drug Administration. Center for Drug Evaluation and Research, 2001.

[15] National Toxicology Program. Toxicology and carcinogenesis studies of 3,3',4,4'-tetrachloroazobenzene. TR-558. (CAS No. 14047-09-7) in Sprague-Dawley rats and B6C3F1 mice (Gavage studies). Technical report, 2010.

[16] National Toxicology Program. Toxicology and carcinogenesis studies of sodium dichromate dihydrate (CAS No. 7789-12-0) in F344/N rats and B6C3F1 mice (Drinking water studies). Technical report, 2010.

[17] National Toxicology Program. Carcinogenicity studies on beta-picoline in F344 rats and B6C3F1 mice. TR-580. Technical report, 2014.

[18] Donner A. and Zou G.Y. Estimating simultaneous confidence intervals for multiple contrasts of proportions by the method of variance estimates recovery. *Statistics in Biopharmaceutical Research, 3:2, 320-335,*, 2011.

[19] O. A. Adaramoye, O. A. Adesanoye, O. M. Adewumi, and O. Akanni. Studies on the toxicological effect of nevirapine, an antiretroviral drug, on the liver, kidney and testis of male Wistar rats. *Human and Experimental Toxicology*, 31(7):676–685, 2012.

[20] I.D. Adler and U. Kliesch. Comparison of single and multiple treatment regimens in the mouse bone-marrow micronucleus assay for hydroquinone and cyclophosphamidecomparison of single and multiple treatment regimens in the mouse bone-marrow micronucleus assay for hydroquinone and cyclophosphamide. *Mutation Reseach*, 234(3-4):115–123, JUN-AUG 1990.

[21] S. Aebtarm and N. Bouguila. An empirical evaluation of attribute control charts for monitoring defects. *Expert Systems with Applications*, 38(6):7869–7880, 2011.

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

[23] A. Agresti and B. A. Coull. Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics*, 52(3):1103–1111, September 1996.

[24] A. Agresti and B. Klingenberg. Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 54:691–706, 2005.

[25] B.C. Allen, R. J. Kavlock, C. A. Kimmel, and E. M. Faustman. Dose-response assessment for developmental toxicity .3. statistical-models. *Fundamental and Applied Toxicology*, 23(4):496–509, November 1994.

[26] D. G. Altman. Why we need confidence intervals. *World Journal of Surgery*, 29(5):554–556, 2005.

[27] D.G. Altman and J.M. Bland. Statistics notes – absence of evidence is not evidence of absence. *British Medical Journal*, 311(7003):485–485, 1995.

[28] D.G. Altman and J.M. Bland. Confidence intervals illuminate absence of evidence. *British Medical Journal*, 328(7446):1016–1017, April 2004.

[29] D.G. Altman and J.M. Bland. Practice statistics notes parametric v non-parametric methods for data analysis. *British Medical Journal*, 338:a3167, April 2009.

[30] H. Andersen, S. Larsen, H. Spliid, and N. D. Christensen. Multivariate statistical analysis of organ weights in toxicity studies. *Toxicology*, 136(2-3):67–77, 1999.

[31] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[32] D.A. Armstrong. factorplot: Improving presentation of simple contrasts in generalized linear models. *R Journal*, 5(2):4–15, 2013.

[33] A.J. Bailer and J.T. Oris. *Assessing Toxicity of Pollutants in Aquatic Systems.* John Wiley, 1994.

[34] A.J. Bailer and C. J. Portier. Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics*, 44(2):417–431, 1988.

[35] J.S. Ball, D.B. Stedman, J. M. Hillegass, C.X. Zhang, J. Panzica-Kelly, and A. et al. Coburn. Fishing for teratogens: A consortium effort for a harmonized zebrafish developmental toxicology assay. *Toxicological Sciences*, 139(1):210–219, May 2014.

[36] D.J. Bartholomew. A test of homogeneity for ordered alternatives. *Biometrika*, 46(1-2):36–48, 1959.

[37] P. Bauer, J. Rohmel, W. Maurer, and L. Hothorn. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*, 17(18):2133–2146, 1998.

[38] F.A. Beland, P.W. Mellick, G.R. Olson, M.C.B. Mendoza, M.M. Marques, and D.R. Doerge. Carcinogenicity of acrylamide in b6c3f(1) mice and f344/n rats from a 2-year drinking water exposure. *Food and Chemical Toxicology*, 51:149–159, 2013.

[39] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 57(1):289–300, 1995.

[40] R. L. Berger and J. C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302, 1996.

[41] G. S. Bieler and R. L. Williams. Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. *Biometrics*, 49(3):793–801, 1993.

[42] E. Biesheuvel. EMA Workshop on Multiplicity Issues in Clinical Trials 16 November 2012, EMA, London, UK. 2012.

[43] E. Billoir, M.L. Delignette-Muller, A.R. R. Pery, and S. Charles. A Bayesian approach to analyzing ecotoxicological data. *Environmental Science & Technology*, 42(23):8978–8984, 2008.

[44] M. B. Black, B. B. Parks, L. Pluta, T. M. Chu, B. C. Allen, R. D. Wolfinger, and R. S. Thomas. Comparison of microarrays and RNA-Seq for gene expression analyses of dose-response experiments. *Toxicological Sciences*, 137(2):385–403, 2014.

[45] J. M. Bland. The tyranny of power: Is there a better way to calculate sample size? *British Medical Journal*, 339:b3985, 2009.

[46] J. M. Bland and D. G. Altman. Practice statistics notes analysis of continuous data from small samples. *British Medical Journal*, 338:a3166, 2009.

[47] E. Bofinger and M. Bofinger. Equivalence with respect to a control: Stepwise tests. *Journal of the Royal Statistical Society B*, 57(4):721–733, 1995.

[48] J. A. Bogoni, N. Armiliato, C. T. Araldi-Favassa, and V. H. Techio. Genotoxicity in Astyanax bimaculatus (twospot astyanax) exposed to the waters of Engano River (Brazil) as determined by micronucleus tests in erythrocytes. *Archives of Environmental Contamination and Toxicology*, 66(3):441–449, 2014.

[49] J. F. Borzelleca. Paracelsus: Herald of modern toxicology. *Toxicological Sciences*, 53(1):2–4, 2000.

[50] W. Brannath and S. Schmidt. A new class of powerful and informative simultaneous confidence intervals. *Statistics in Medicine*, 33(19):3365–3386, 2014.

[51] F. Bretz. An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics and Data Analysis*, 50(7):1735–1748, 2006.

[52] F. Bretz and L. A. Hothorn. Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. *Statistics in Medicine*, 21(22):3325–3335, 2002.

[53] F. Bretz, L. A. Hothorn, and J. C. Hsu. Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Statistics in Medicine*, 22(6):847–858, 2003.

[54] F. Bretz and L.A. Hothorn. Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. *ATLA-Alternatives to Laboratory Animals*, 31(Suppl. 1):81–96, 2003.

[55] F. Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. Chapman and Hall/CRC, 0 edition, 7 2010.

[56] F. Bretz and D. Seidel. Sas/iml programs for calculating orthant probabilities for arbitrary dimensions. *Computational Statistics & Data Analysis*, 33(2):217–218, 2000.

[57] J. Bright, M. Aylott, S. Bate, H. Geys, P. Jarvis, J. Saul, and R. Vonk. Recommendations on the statistical analysis of the Comet assay. *Pharmaceutical Statistics*, 10(6, SI):485–493, 2011.

[58] I. D. Bross. Why proof of safety is much more difficult than proof of hazard. *Biometrics*, 41(3):785–793, 1985.

[59] C. C. Brown and T. R. Fears. Exact significance levels for multiple binomial testing with application to carcinogenicity screens. *Biometrics*, 37(4):763–774, 1981.

[60] L. D. Brown, T. T. Cai, A. DasGupta, A. Agresti, B. A. Coull, and G. et al. Casella. Interval estimation for a binomial proportion - comment - rejoinder. *Statistical Science*, 16(2):101–133, 2001.

[61] R. H. Browne. The t-test p value and its relationship to the effect size and p(x > y). *American Statistician*, 64(1):30–33, 2010.

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

[63] E. Brunner and Munzel U. *Nichtparametrische Datenanalyse. Unverbundene Stichproben. Statistik und ihre Anwendungen.* Springer Heidelberg, 2002.

[64] R. Buesen, R. Landsiedel, U. G. Sauer, W. Wohlleben, S. Groeters, V. Strauss, H. Kamp, and B. van Ravenzwaay. Effects of SiO2, ZrO2, and BaSO4 nanomaterials with or without surface functionalization upon 28-day oral exposure to rats. *Archives of Toxicology*, 88(10):1881–1906, 2014.

[65] H. Buning. Robust and adaptive tests for the 2-sample location problem. *Operational Research*, 16(1):33–39, 1994.

[66] H.U. Burger, U. Beyer, and M. Abt. Issues in the assessment of non-inferiority: Perspectives drawn from case studies. *Pharmaceutical Statistics*, 10(5):433–439, 2011.

[67] F. Cabanne, J. C. Gaudry, and J. C. Streibig. Influence of alkyl oleates on efficacy of phenmedipham applied as an acetone: Water solution on galium aparine. *Weed Research*, 39(1):57–67, 1999.

[68] I. Campbell. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26(19):3661–3675, 2007.

[69] A. Canty and B.D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2014. R package version 1.3-11.

[70] N. F. Cariello and W. W. Piegorsch. The Ames test: The two-fold rule revisited. *Mutation Research-Genetic Toxicology*, 369(1-2):23–31, 1996.

[71] K. C. Carriere. How good is a normal approximation for rates and proportions of low incidence events? *Communications in Statistics-Simulation and Computation*, 30(2):327–337, 2001.

[72] P. J. Catalano and L. M. Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419):651–658, 1992.

[73] P. J. Catalano, D. O. Scharfstein, and L. Ryan. Statistical-model for fetal death, fetal weight, and malformation in developmental toxicity studies. *Teratology*, 47(4):281–290, 1993.

[74] Y. H. Chen and X. H. Zhou. Interval estimates for the ratio and difference of two lognormal means. *Statistics in Medicine*, 25(23):4099–4113, December 2006.

[75] Y. J. Chen, K. C. Lai, H. H. Kuo, L. P. Chow, L. H. Yih, and T. C. Lee. HSP70 colocalizes with PLK1 at the centrosome and disturbs spindle dynamics in cells arrested in mitosis by arsenic trioxide. *Archives of Toxicology*, 88(9):1711–1723, 2014.

[76] Z. Chen, B. Zhang, and P. S. Albert. A joint modeling approach to data with informative cluster size: Robustness to the cluster size model. *Statistics in Medicine*, 30(15):1825–1836, 2011.

[77] R. H. B. Christensen. ordinal—regression models for ordinal data, 2015. R package version 2015.1-21. http://www.cran.r-project.org/package=ordinal/.

[78] C. Clark, C. Schreiner, C. Parker, T. Gray, and G.M. Hoffman. Health assessment of gasoline and fuel oxygenate vapors: Subchronic inhalation toxicity. *Regulatory Toxicology and Pharmacology*, 70(2):S18–S28, 2014.

[79] W. J. Conover and R. L. Iman. Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, 35(3):124–129, 1981.

[80] W. J. Conover and D. S. Salsburg. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to respond to treatment. *Biometrics*, 44(1):189–196, 1988.

[81] J. D. Consiglio, G. Shan, and G. E. Wilding. A comparison of exact tests for trend with binary endpoints using Bartholomew's statistic. *International Journal of Biostatistics*, 10(2):221–230, 2014.

[82] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. B*, 34(2):187–220, 1972.

[83] G. G. Crans and J. J. Shuster. How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. *Statistics in Medicine*, 27(18):3598–3611, 2008.

[84] D. Curran-Everett. Explorations in statistics: The analysis of ratios and normalized data. *Advances in Physiology Education*, 37(3):213–219, 2013.

[85] L.L. Curry and A. Roberts. Subchronic toxicity of rebaudioside A. *Food and Chemical Toxicology 46 (2008) S11*, 46:S11–S20, 2008.

[86] D. B. Dahl. *xtable: Export tables to LaTeX or HTML*, 2014. R package version 1.7-3.

[87] O. Davidov and S. Peddada. Order-restricted inference for multivariate binary data with application to toxicology. *Journal of the American Statistical Association*, 106(496):1394–1404, 2011.

[88] O. Davidov and S. Peddada. Testing for the multivariate stochastic order among ordered experimental groups with application to dose-response studies. *Biometrics*, 69(4):982–990, 2013.

[89] H. D. Delaney and A. Vargha. Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, 7(4):485–503, 2002.

[90] M.L. Delignette-Muller, C. Forfait, E. Billoir, and S. Charles. A new perspective of the Dunnett procedure: filling the gap between NOEC/LOEC and EXx concepts. *Environmental Toxicology and Chemistry*, 30(12):2888–2891, 2011.

[91] M.L. Delignette-Muller, P. Ruiz, S. Charles, W. Duchemin, C. Lopes, and V. Veber. *morse: MOdelling tools for Reproduction and Survival Data in Ecotoxicology*, 2014. R package version 1.0.2.

[92] D. L. Denton, J. Diamond, and L. Zheng. Test of significance in toxicity: A statistical application for assessing whether an effluent or site water is truly toxic. *Environmental Toxicology and Chemistry*, 30(5):1117–1126, 2011.

[93] H. Dette and A. Munk. Optimum allocation of treatments for welch's test in equivalence assessment. *Biometrics*, 53(3):1143–1150, 1997.

[94] J. M. Diamond, D. L. Denton, J. W. Roberts, and L. Zheng. Evaluation of the test of significant toxicity for determining the toxicity of effluents and ambient water samples. *Environmental Toxicology and Chemistry*, 32(5):1101–1108, 2013.

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

[96] G. Dilba, F. Bretz, L. A. Hothorn, and V. Guiard. Power and sample size computations in simultaneous tests for non-inferiority based on relative margins. *Statistics in Medicine*, 25(7):1131–1147, 2006.

[97] G. Dilba, F. Schaarschmidt, and L.A. Hothorn. Inferences for ratios of normal means. *R News*, 7:20–23, 2007.

[98] R. P. Do, R. W. Stahlhut, D. Ponzi, F. S. vom Saal, and J. A. Taylor. Non-monotonic dose effects of in utero exposure to di(2-ethylhexyl) phthalate (dehp) on testicular and serum testosterone and anogenital distance in male mouse fetuses. *Reproductive Toxicology*, 34(4):614–621, 2012.

[99] J. B. du Prel, G. Hommel, B. Rohrig, and M. Blettner. Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International*, 106(19):335–339, 2009.

[100] J. B. du Prel, B. Rohrig, G. Hommel, and M. Blettner. Choosing statistical tests. *Deutsches Arzteblatt International*, 107(19):343–348, 2010.

[101] O. J. Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–&, 1964.

[102] 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.

[103] D. B. Dunson, Z. Chen, and J. Harry. A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*, 59(3):521–530, 2003.

[104] G. Ehling, M. Hecht, A. Heusener, J. Huesler, A. O. Gamer, H. van Loveren, T. Maurer, K. Riecke, L. Ullmann, P. Ulrich, R. Vandebriel, and H. W. Vohr. An european interlaboratory validation of alternative endpoints of the murine local lymph node assay - 2nd round. *Toxicology*, 212(1):69–79, 2005.

[105] L. Einaudi, B. Courbiere, V. Tassistro, C. Prevot, I. Sari-Minodier, T. Orsiere, and J. Perrin. In vivo exposure to benzo(a) pyrene induces significant DNA damage in mouse oocytes and cumulus cells. *Human Reproduction*, 29(3):548–554, 2014.

[106] A. Ejchart and N. Sadlej-Sosnowska. Statistical evaluation and comparison of comet assay results. *Mutation Research-Genetic Toxicology and Environmental Mutagenesis*, 534(1-2):85–92, 2003.

[107] M. R. Elliott, M. M. Joffe, and Z. Chen. A potential outcomes approach to developmental toxicity analyses. *Biometrics*, 62(2):352–360, 2006.

[108] S. A. Elmore and S. D. Peddada. Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. *Toxicologic Pathology*, 37:672–676, 2009.

[109] M. Elwell, W. Fairweather, X. Fouillet, K. Keenan, K. Lin, G. Long, L. Mixson, D. Morton, T. Peters, C. Rousseaux, and D. Tuomari. The society of toxicologic pathology's recommendations on statistical analysis of rodent carcinogenicity studies. *Toxicologic Pathology*, 30(3):415–418, 2002.

[110] G. Engelhardt. In vivo micronucleus test in mice with 1-phenylethanol. *Archives of Toxicology*, 80(12):868–872, 2006.

[111] W. P. Erickson and L. L. McDonald. Tests for bioequivalence of control media and test media in studies of toxicity. *Environmental Toxicology and Chemistry*, 14(7):1247–1256, 1995.

[112] C. Eskes, S. Hoffmann, D. Facchini, R. Ulmer, A. Wang, M. Flego, M. Vassallo, M. Bufo, E. van Vliet, F. d'Abrosca, and N. Wilt. Validation study on the ocular irritection (r) assay for eye irritation testing. *Toxicology in Vitro*, 28(5):1046–1065, 2014.

[113] C. Faes, M. Aerts, H. Geys, and L. De Schaepdrijver. Modeling spatial learning in rats based on Morris water maze experiments. *Pharmaceutical Statistics*, 9(1):10–20, 2010.

[114] C. Faes, M. Aerts, H. Geys, G. Molenberghs, and L. Declerck. Bayesian testing for trend in a power model for clustered binary data. *Environmental and Ecological Statistics*, 11(3):305–322, 2004.

[115] M. F. W. Festing. Extending the statistical analysis and graphical presentation of toxicity test results using standardized effect sizes. *Toxicologic Pathology*, 42(8):1238–1249, 2014.

[116] E. C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 16(2):175–185, 1954.

[117] P.M. Forster. Explanation of levels of evidence for developmental toxicity. Technical report, US-NTP (http://ntp.niehs.nih.gov/go/10003), 2014.

[118] M. F. Freeman and J. W. Tukey. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, 21(4):607–611, 1950.

[119] A. O. Gamer, R. Rossbacher, W. Kaufmann, and B. van Ravenzwaay. The inhalation toxicity of di- and triethanolamine upon repeated exposure. *Food and Chemical Toxicology*, 46(6):2173–2183, 2008.

[120] J. J. Gart and J. Nam. Approximate interval estimation of the ratio of binomial parameters - a review and corrections for skewness. *Biometrics*, 44(2):323–338, 1988.

[121] Dilba G.D., M. Hasler, D. Gerhard, and F. Schaarschmidt. *mratios: Inferences for ratios of coefficients in the general linear model*, 2012. R package version 1.3.17.

[122] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

[123] A. Genz and F. Bretz. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4):361–378, 1999.

[124] D. Gerhard. Simultaneous small sample inference for linear combinations of generalized linear model parameters. *Communications in Statistics - Simulation and Computation*, 2014.

[125] D. Gerhard, L.A. Hothorn, and R. Vonk. Statistical evaluation of the in vivo comet assay taking biological relevance into account. Technical report, Reports of the Institute of Biostatistics No 10 / 2008 Leibniz University of Hannover Natural Sciences Faculty, 2008.

[126] D. Gerhard and R.M. Kuiper. *goric: Generalized Order-Restricted Information Criterion*, 2014. R package version 0.0-8.

[127] D Gerhard and F. Schaarschmidt. *mmcp: Multiple Comparison Procedures for Multinomial Models*, 2014. R package version 0.0-8.

[128] D. Gerhard and F. Schaarschmidt. *mmcp: Multiple Comparison Procedures for Multinomial Models*, 2014. R package version 0.0-8.

[129] A. K. Goetz, B. P. Singh, M. Battalora, J. M. Breier, J. P. Bailey, A. C. Chukwudebe, and E. R. Janus. Current and future use of genomics data in toxicology: Opportunities and challenges for regulatory applications. *Regulatory Toxicology and Pharmacology*, 61(2):141–153, 2011.

[130] B. I. Graubard and E. L. Korn. Choice of column scores for testing independence in ordered 2 X K contingency-tables. *Biometrics*, 43(2):471–476, 1987.

[131] J. Green and J. R. Wheeler. The use of carrier solvents in regulatory aquatic toxicology testing: Practical, statistical and regulatory considerations. *Aquatic Toxicology*, 144:242–249, 2013.

[132] J. W. Green. Power and control choice in aquatic experiments with solvents. *Ecotoxicology and Environmental Safety*, 102:142–146, April 2014.

[133] J. W. Green, T. A. Springer, A. N. Saulnier, and J. Swintek. Statistical analysis of histopathological endpoints. *Environmental Toxicology and Chemistry*, 33(5):1108–1116, 2014.

[134] B. Grün and F. Leisch. FlexMix: An R package for finite mixture modelling. *R News*, 7(1):8–13, April 2007.

[135] Y. Guan. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. *Statistics & Probability Letters*, 79(14):1621–1629, 2009.

[136] R. V. Gueorguieva. Comments about joint modeling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics*, 61(3):862–866, 2005.

[137] R. V. Gueorguieva and G. Sanacora. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, 25(8):1307–1322, 2006.

[138] M. J. Gurka. Selecting the best linear mixed model under reml. *American Statistician*, 60(1):19–26, 2006.

[139] Goehlmann H. and W. Talloen. *Gene Expression Studies Using Affymetrix Microarrays*. Chapman and Hall, 2009.

[140] G. Hahn and W.Q. Meeker. *Statistical Intervals – A Guide for Practitioners*. John Wiley and Sons, Inc., New York, 1991.

[141] M. Hardy and T. Stedeford. Developmental neurotoxicity: When research succeeds through inappropriate statistics. *Neurotoxicology*, 29(3):476–476, 2008.

[142] M. Hardy and T. Stedeford. Use of the pup as the statistical unit in developmental neurotoxicity studies: Overlooked model or poor research design? *Toxicological Sciences*, 103(2):409–410, 2008.

[143] M. Hasler. Multiple comparisons to both a negative and a positive control. *Pharmaceutical Statistics*, 11(1):74–81, 2012.

[144] M. Hasler and L. A. Hothorn. A multivariate Williams-type trend procedure. *Statistics in Biopharmaceutical Research*, 4(1):57–65, 2012.

[145] M. Hasler and L. A. Hothorn. Simultaneous confidence intervals on multivariate non-inferiority. *Statistics in Medicine*, 32(10):1720–1729, 2013.

[146] M. Hasler and L.A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 51:1, 2008.

[147] M. Hasler, R. Vonk, and L. A. Hothorn. Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity. *Statistics in Medicine*, 27(4):490–503, 2008.

[148] Mario Hasler. *ETC: Equivalence to control*, 2009. R package version 1.3.

[149] Mario Hasler. *SimComp: Simultaneous Comparisons for Multiple Endpoints*, 2014. R package version 2.2.

[150] D. Hauschke, T. Hothorn, and J. Schafer. The role of control groups in mutagenicity studies: Matching biological and statistical relevance. *ATLA-Alternatives to Laboratory Animals*, 31:65–75, June 2003.

[151] D. Hauschke, M. Kieser, and L. A. Hothorn. Proof of safety in toxicology based on the ratio of two means for normally distributed data. *Biometrical Journal*, 41(3):295–304, 1999.

[152] D. Hauschke, R. Slacik-Erben, S. Hensen, and R. Kaufmann. Biostatistical assessment of mutagenicity studies by including the positive control. *Biometrical Journal*, 47(1):82–87, 2005.

[153] M. Hayashi, K. Dearfield, P. Kasper, D. Lovell, H.J. Martus, and V. Thybaud. Compilation and use of genetic toxicity historical control data. *Mutation Research-Genetic Toxicology and Environmental Mutagenesis*, 723(2):87–90, 2011.

[154] A. F. Hayes and L. Cai. Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical & Statistical Psychology*, 60:217–244, 2007.

[155] A. J. Hayter. Inferences on the difference between future observations for comparing two treatments. *Journal of Applied Statistics*, 40(4):887–900, 2013.

[156] J. L. He, W. L. Chen, L. F. Jin, and H. Y. Jin. Comparative evaluation of the in vitro micronucleus test and the comet assay for the detection of genotoxic effects of x-ray radiation. *Mutation Research-Genetic Toxicology and Environmental Mutagenesis*, 469(2):223–231, 2000.

[157] E. Herberich and L.A. Hothorn. Statistical evaluation of mortality in long-term carcinogenicity bioassays using a Williams-type procedure. *Regulatory Toxicology and Pharmacology*, 64:26–34, 2012.

[158] E. Herberich and T. Hothorn. Dunnett-type inference in the frailty Cox model with covariates. *Statistics in Medicine*, 31(1):45–55, 2012.

[159] E. Herberich, J. Sikorski, and T. Hothorn. A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PLOS One*, 5(3):e9788, 2010.

[160] C. Hirotsu, S. Yamamoto, and L.A. Hothorn. Estimating the dose-response pattern by the maximal contrast type test approach. *Statistics in Biopharmaceutical Research*, 3(1):40–53, 2011.

[161] S. Højsgaard, U. Halekoh, and Jun Y. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15/2:1–11, 2006.

[162] A. M. Hoberman, D. K. Schreur, T. Leazer, G. P. Daston, P. Carthew, T. Re, L. Loretz, and P. Mann. Lack of effect of butylparaben and methylparaben on the reproductive system in male rats. *Birth Defects Research Part B-Developmental and Reproductive Toxicology*, 83(2):123–133, 2008.

[163] J. L. Hodges and E. L. Lehmann. Estimates of location based on rank-tests. *Annals of Mathematical Statistics*, 34(2):598–&, 1963.

[164] W. P. Hoffman, D. K. Ness, and R. B. L. van Lier. Analysis of rodent growth data in toxicology studies. *Toxicological Sciences*, 66(2):313–319, 2002.

[165] W. P. Hoffman, J. Recknor, and C. Lee. Overall type I error rate and power of multiple Dunnett's tests on rodent body weights in toxicology studies. *Journal of Biopharmaceutical Statistics*, 18(5):883–900, 2008.

[166] S. Hoffmann, L. A. Hothorn, L. Edler, A. Kleensang, M. Suzuki, P. Phrakonkham, and D. Gerhard. Two new approaches to improve the analysis of BALB/c 3T3 cell transformation assay data. *Mutation Research-Genetic Toxicology and Environmental Mutagenesis*, 744(1):36–41, 2012.

[167] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[168] P. M. Hooper and Z. L. Yang. Confidence intervals following Box-Cox transformation. *Canadian Journal of Statistics-revue Canadienne De Statistique*, 25(3):401–416, 1997.

[169] L. Hothorn. Robustness study on Williams procedure and Shirley procedure, with application In toxicology. *Biometrical Journal*, 31(8):891–903, 1989.

[170] L. Hothorn. Biostatistical analysis of the micronucleus mutagenicity assay based on the sssumption of a mixing distribution. *Environmental Health Perspectives*, 102:121–125, 1994.

[171] L. Hothorn and W. Lehmacher. A simple testing procedure control versus k treatments for one-sided ordered-alternatives, with application in toxicology. *Biometrical Journal*, 33(2):179–189, 1991.

[172] L. A. Hothorn. Statistics of interlaboratory in vitro toxicological studies. *ATLA-Alternatives to Laboratory Animals*, 31:43–63, 2003.

[173] L. A. Hothorn. Multiple comparisons and multiple contrasts in randomized dose-response trials-confidence interval oriented approaches. *Journal of Biopharmaceutical Statistics*, 16(5):711–731, 2006.

[174] L. A. Hothorn and F. Bretz. Evaluation of animal carcinogenicity studies: Cochran-Armitage trend test vs. multiple contrast tests. *Biometrical Journal*, 42(5):553–567, 2000.

[175] L. A. Hothorn and F. Bretz. Dose-response and thresholds in mutagenicity studies: A statistical testing approach. *ATLA-Alternatives to Laboratory Animals*, 31:97–103, 2003.

[176] L. A. Hothorn and D. Gerhard. Statistical evaluation of the *in vivo* micronucleus assay. *Archives of Toxicology*, 83(6):625–634, 2009.

[177] L. A. Hothorn and T. Hothorn. Order-restricted scores test for the evaluation of population-based case-control studies when the genetic model is unknown. *Biometrical Journal*, 51(4):659–669, 2009.

[178] 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.

[179] L. A. Hothorn, K. Reisinger, T. Wolf, A. Poth, D. Fieblingere, M. Liebsch, and R. Pirow. Statistical analysis of the hen's egg test for micronucleus induction (het-mn assay). *Mutation Research-Genetic Toxicology and Environmental Mutagenesis*, 757(1):68–78, 2013.

[180] L. A. Hothorn and H.W. Vohr. Statistical evaluation of the Local Lymph Node Assay. *Regulatory Toxicology and Pharmacology*, 56(3):352–356, 2010.

[181] L.A. Hothorn. *Regulatory Toxicology*, chapter tatistical Evaluation Methods in Toxicology, pages 213–223. Springer Heidelberg, 2014.

[182] L.A. Hothorn. *SiTuR: Data files for Statistics in Toxicology using R*, 2014. R package version 1.0.

[183] L.A. Hothorn. The two-step approach - a significant ANOVA F-test before Dunnett's comparisons against a control - is not recommended. *Communications in Statistics*, 2015.

[184] L.A. Hothorn and G.D. Dilba. A ratio-to-control Williams-type test for trend. *Pharmaceutical Statistics*, 11:1111, 2010.

[185] L.A. Hothorn and M. Hasler. Proof of hazard and proof of safety in toxicological studies using simultaneous confidence intervals for differences and ratios to control. *Journal of Biopharmaceutical Statistics*, 18:915–933, 2008.

[186] L.A. Hothorn and D. Hauschke. Identifying the maximum safe dose: a multiple testing approach. *Journal of Biopharmaceutical Statistics 10: 15-30.*, 2000.

[187] L.A. Hothorn and F. Schaarschmidt. One-sided ratio-to-control-tests - simulation results. Technical report, Leibniz University Hannover, Institute of Biostatistics, 2014.

[188] L.A. Hothorn, M. Sill, and F. Schaarschmidt. Evaluation of incidence rates in pre-clinical studies using a Williams-type procedure. *The International Journal of Biostatistics*, 6:15, 2010.

[189] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

[190] T. Hothorn, F. Bretz, and P. Westfall. *multcomp: Simultaneous Inference for General Linear Hypotheses.*, 2011. R package version 1.2-6, <http://CRAN.R-project.org/package=multcomp>.

[191] T. Hothorn, L. Held, and T. Friede. Biometrical journal and reproducible research. *Biometrical Journal*, 51(4):553–555, 2009.

[192] T. Hothorn, K. Hornik, M. van de Wiel, and A. Zeileis. A lego system for conditional inference. *The American Statistician*, 60(3):257–263, 2006.

[193] T Hothorn, K Hornik, MA van de Wiel, and A Zeileis. *coin: Conditional Inference Procedures in a Permutation Test Framework*, 2007. Package version 0.5-2.

[194] T. Hothorn, K. Hornik, M.A. van de Wiel, and A. Zeileis. Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28(8):1–23, 2008.

[195] J. F. Howell and P. A. Games. Effects of variance heterogeneity on simultaneous multiple comparison procedures with equal sample size. *British Journal of Mathematical & Statistical Psychology*, 27(5):72–81, 1974.

[196] D. L. Hunt, S. N. Rai, and C. S. Li. Summary of dose-response modeling for developmental toxicity studies. *Dose-Response*, 6(4):352–368, 2008.

[197] B. S. Hwang and M. L. Pennell. Semiparametric Bayesian joint modeling of a binary and continuous outcome with applications in toxicological risk assessment. *Statistics in Medicine*, 33(7):1162–1175, 2014.

[198] ICH-E9. Statistical principles for clinical trials. Technical report, CPMP/ICH/363, 1998.

[199] ICH-S2. Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use. Technical report, ICH, 2008.

[200] ICH-S5A. Reproductive toxicology: Detection of toxicity to reproduction for medicinal products including toxicity to male fertility. Technical report, CPMP/ICH/386/95, 1994.

[201] National Cancer Institute. Carcinogenesis technical report series no. 142 1979 bioassay of p-cresidine for possible carcinogenicity nci-cg-tr-142. Technical report, 1979.

[202] R. J. Isfort, G. A. Kerckaert, and R. A. LeBoeuf. Comparison of the standard and reduced ph Syrian hamster embryo (she) cell *in vitro* transformation assays in predicting the carcinogenic potential of chemicals. *Mutation Research- Fundamental and Molecular Mechanisms of Mutagenesis*, 356(1):11–63, 1996.

[203] C.C. Jacob, R. Reimschuessel, and et al. vonTungeln, L.S. Dose-response assessment of nephrotoxicity from a 7-day combined exposure to melamine and cyanuric acid in f344 rats. *Toxicological Sciences*, 119(2):391–397, 2011.

[204] T. Jaki and L. A. Hothorn. Statistical evaluation of toxicological assays: Dunnett or williams test-take both. *Archives of Toxicology*, 87(11):1901–1910, 2013.

[205] T. Jaki and M. J. Wolfsegger. Non-compartmental estimation of pharmacokinetic parameters for flexible sampling designs. *Statistics in Medicine*, 31(11-12):1059–1073, 2012.

[206] Thomas Jaki and Martin Wolfsegger. Estimation of pharmacokinetic parameters with the r package pk. *Pharmaceutical Statistics*, 10(3):284–288, 2011. DOI: 10.1002/pst.449.

[207] P. Jarvis, J. Saul, M. Aylott, S. Bate, H. Geys, and J. Sherington. An assessment of the statistical methods used to analyse toxicology studies. *Pharmaceutical Statistics*, 10(6, SI):477–484, 2011.

[208] P. W. Jones, K. M. Beeh, K. R. Chapman, M. Decramer, D. A. Mahler, and J. A. Wedzicha. Minimal clinically important differences in pharmacological trials. *American Journal of Respiratory and Critical Care Medicine*, 189(3):250–255, 2014.

[209] J. Kanno, L. Onyon, S. Peddada, J. Ashby, E. Jacob, and W. Owens. The OECD program to validate the rat uterotrophic bioassay. phase 2: Dose-response studies. *Environmental Health Perspectives*, 111(12):1530–1549, 2003.

[210] E. Kasuya. Mann-Whitney U test when variances are unequal. *Animal Behaviour*, 61:1247–1249, June 2001.

[211] G. A. Kerckaert, R. Brauninger, R. A. LeBoeuf, and R. J. Isfort. Use of the Syrian hamster embryo cell transformation assay for carcinogenicity prediction of chemicals currently being tested by the national toxicology program in rodent bioassays. *Environmental Health Perspectives*, 104:1075–1084, 1996.

[212] M. Kieser, T. Friede, and M. Gondan. Assessment of statistical significance and clinical relevance. *Statistics in Medicine*, 32(10):1707–1719, 2013.

[213] B. S. Kim, M. H. Cho, and H. J. Kim. Statistical analysis of in vivo rodent micronucleus assay. *Mutation Research-Genetic Toxicology And Environmental Mutagenesis*, 469(2):233–241, 2000.

[214] A. Kitsche and L. A. Hothorn. Testing for qualitative interaction using ratios of treatment differences. *Statistics in Medicine*, 33(9):1477–1489, 2014.

[215] A. Kitsche, L. A. Hothorn, and F. Schaarschmidt. The use of historical controls in estimation simultaneous confidence intervals for comparisons against a concurrent control. *Computational Statistics and Data Analysis*, 56(12):3865–3875, 2012.

[216] A. Kitsche and F. Schaarschmidt. Analysis of statistical interactions in factorial experiments. *Journal of Agronomy and Crop Science*, 2014.

[217] B. Klingenberg. A new and improved confidence interval for the Mantel-Haenszel risk difference. *Statistics in Medicine*, 33(17):2968–2983, 2014.

[218] B. Klingenberg and V. Satopaa. Simultaneous confidence intervals for comparing margins of multivariate binary data. *Computational Statistics & Data Analysis*, 64:87–98, 2013.

[219] K. Kobayashi, K. Pillai, M. Michael, K.M. Cherian, A. Araki, and A. Hirose. Determination of dose dependence in repeated dose toxicity studies when mid-dose alone is insignificant. *Journal of Toxicological Sciences*, 37(2):255–260, 2012.

[220] K. Kobayashi, Y. Sakuratani, T. Abe, S. Nishikawa, J. Yamada, A. Hirose, E. Kamata, and M. Hayashi. Relation between statistics and treatment-related changes obtained from toxicity studies in rats: if detected a significant difference in low or middle dose for quantitative values, this change is considered as incidental change? *Journal of Toxicological Sciences*, 35(1):79–85, February 2010.

[221] H. F. Koch and L. A. Hothorn. Exact unconditional distributions for dichotomous data in many-to-one comparisons. *Journal of Statistical Planning and Inference*, 82(1-2):83–99, 1999.

[222] R. L. Kodell. Should we assess tumorigenicity with the Peto or Poly-k test? *Statistics in Biopharmaceutical Research*, 4(2):118–124, 2012.

[223] R.L. Kodell. Replace the NOAEL and LOAEL with the BMDL01 and BMDL10. *Environmental and Ecological Statistics*, 16(1):3–12, 2009.

[224] J. E. Kolassa. A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical-data. *Statistics in Medicine*, 14(14):1577–1581, 1995.

[225] F. Konietschke. *Simultane Konfidenzintervalle fuer nichtparametrische relative Kontrasteffekte*. PhD thesis, Georg-August-Universitaet Goettingen, 2009.

[226] F. Konietschke, A. C. Batke, L. A. Hothorn, and E. Brunner. Testing and estimation of purely nonparametric effects in repeated measures designs. *Computational Statistics & Data Analysis*, 54(8):1895–1905, 2010.

[227] F. Konietschke, S. Bosiger, E. Brunner, and L. A. Hothorn. Are multiple contrast tests superior to the anova? *The International Journal of Biostatistics*, 9(1), 2013.

[228] F. Konietschke and L.A. Hothorn. Evaluation of toxicological studies using a non-parametric Shirley-type trend test for comparing several dose levels with a control group. *Statistics in Biopharmaceutical Research*, 4:14–27, 2012.

[229] F. Konietschke and L.A. Hothorn. Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, 6:738–759, 2012.

[230] F. Konietschke, M. Placzek, F. Schaarschmidt, and L.A. Hothorn. nparcomp: An R software package for nonparametric multiple comparisons and simultaneous confidence intervals. *Journal of Statistical Software*, 64(9):1–17, 2015.

[231] A. K. Krug, R. Kolde, J. A. Gaspar, E. Rempel, N. V. Balmer, K. Meganathan, K. Vojnits, M. Baquie, T. Waldmann, R. Ensenat-Waser, S. Jagtap, R. M. Evans, S. Julien, H. Peterson, D. Zagoura, S. Kadereit, D. Gerhard, I. Sotiriadou, M. Heke, K. Natarajan, M. Henry, J. Winkler, R. Marchan, L. Stoppini, S. Bosgra, J. Westerhout, M. Verwei, J. Vilo, A. Kortenkamp, J. R. Hescheler, L. Hothorn, S. Bremer, C. van Thriel, K. H. Krause, J. G. Hengstler, J. Rahnenfuhrer, M. Leist, and A. Sachinidis. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Archives of Toxicology*, 87(1):123–143, 2013.

[232] R. M. Kuiper, D. Gerhard, and L. A. Hothorn. Identification of the minimum effective dose for normally distributed endpoints using a model selection approach. *Statistics in Biopharmaceutical Research*, 6(1):55–66, 2014.

[233] L.L. Laster and M.F. Johnson. Non-inferiority trials: the 'at least as good as' criterion. *Statistics in Medicine*, 22(2):187–200, 2003.

[234] R. Lawson. Small sample confidence intervals for the odds ratio. *Communications in Statistics-Simulation and Computation*, 33(4):1095–1113, 2004.

[235] R. A. LeBoeuf, G. A. Kerckaert, M. J. Aardema, D. P. Gibson, R. Brauninger, and R. J. Isfort. The pH 6.7 Syrian hamster embryo cell transformation assay for assessing the carcinogenic potential of chemicals. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis*, 356(1):85–127, 1996.

[236] W. Leisenring and L. Ryan. Statistical properties of the noael. *Regulatory Toxicology and Pharmacology*, 15(2):161–171, April 1992.

[237] Lesnoff, M., Lancelot, and R. *aod: Analysis of Overdispersed Data*, 2012. R package version 1.3.

[238] K. Leuraud and J. Benichou. A comparison of stratified and adjusted trend tests for binomial proportions. *Statistics in Medicine*, 25(3):529–535, 2006.

[239] G. Lewin and T Tilmann. Einfluss niederfrequenter elektromagnetischer felder auf das sich entwickelnde blutbildende system, das immunsystem und das zns in vivo - vorhaben 3608s30006 band 1: Hauptbericht. Technical report, Bundesamt für Strahlenschutz, 2013.

[240] R. W. Lewis, R. Billington, E. Debryune, A. Gamer, B. Lang, and F. Carpanini. Recognition of adverse and nonadverse effects in toxicity studies. *Toxicologic Pathology*, 30(1):66–74, 2002.

[241] D. Lin, S. Pramana, T. Verbeke, and Z. Shkedy. *IsoGene: Order-Restricted Inference for Microarray Experiments*, 2014. R package version 1.0-23.

[242] D. Lin, Z.. Shkedy, and Burzykowski T. Yekutieli, D., H.W. H. Goehlmann, A. De Bondt, T. Perera, T. Geerts, and L. Bijnens. Testing for trends in dose-response microarray experiments: A comparison of several testing procedures, multiplicity and resampling-based inference. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

[243] D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga, and L. Bijnens, editors. *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*. Springer, 2012.

[244] L. J. Lin, D. Bandyopadhyay, S. R. Lipsitz, and D. Sinha. Association models for clustered data with binary and continuous responses. *Biometrics*, 66(1):287–293, 2010.

[245] C.M. Lombardi and S. H. Hurlbert. Misprescription and misuse of one-tailed tests. *Australian Ecology*, 34(4):447–468, 2009.

[246] D. P. Lovell. The use of statistical and quantitative bioinformatic methods in toxicogenomics. *Toxicology*, 240(3):160–161, 2007.

[247] D. P. Lovell and T. Omori. Statistical issues in the use of the comet assay. *Mutagenesis*, 23:1–12, 2008.

[248] S. Lydersen, M. W. Fagerland, and P. Laake. Recommended tests for association in 2 x 2 tables. *Statistics in Medicine*, 28(7):1159–1175, 2009.

[249] R. Manar, P. Vasseur, and H. Bessi. Chronic toxicity of chlordane to Daphnia magna and Ceriodaphnia dubia: A comparative study. *Environmental Toxicology*, 27(2):90–97, 2012.

[250] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Canver Chemotherapy Reports 50: (3) 163-170*, 1966.

[251] N. Mantel, N. R. Bohidar, and J. L. Ciminera. Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information. *Cancer Research*, 37(11):3863–3868, 1977.

[252] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

[253] 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.

[254] R. Marcus and H. Talpaz. Further results on testing homogeneity of normal means against simple tree alternatives. *Communications in Statistics- Theory and Methods*, 21(8):2135–2149, 1992.

[255] B. H. Margolin, N. Kaplan, and E. Zeiger. Statistical analysis of the Ames Salmonella-Microsome test. *Proceedings of the National Academy of Sciences- Biological Sciences*, 78(6):3779–3783, 1981.

[256] C. G. Markgraf, M. Cirino, and J. Meredith. Comparison of methods for analysis of functional observation battery (fob) data. *Journal of Pharmacological and Toxicological Methods*, 62(2):89–94, 2010.

[257] K. Maruo and N. Kawai. Confidence intervals based on some weighting functions for the difference of two binomial proportions. *Statistics in Medicine*, 33(13):2288–2296, 2014.

[258] W. Maurer and and Lehmacher W. Hothorn, L. A. *Biometrie in der chemisch-pharmazeutischen Industrie, Volume 6 (1995)*, chapter Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses, pages 3–18. Fischer Verlag Stuttgart, 1995.

[259] D. V. Mehrotra, I. S. F. Chan, and R. L. Berger. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics*, 59(2):441–450, 2003.

[260] E.J. Meiman. *Effects on Pinniped Immune Response Upon in vitro Exposure to the Perfluorinated Compounds, PFOS and PFOA*. PhD thesis, University of Connnecticut. Honors Scholar Theses. Paper 362., 2014.

[261] B. Michael, B. Yano, R. S. Sellers, R. Perry, D. Morton, N. Roome, J. K. Johnson, and K. Schafer. Evaluation of organ weights for rodent and non-rodent toxicity studies: A review of regulatory guidelines and a survey of current practices. *Toxicologic Pathology*, 35(5):742–750, 2007.

[262] S.P. Millard. Proof of safety vs proof of hazard. *Biometrics*, 43(3):719–725, 1987.

[263] F. K. Mohammad and S. Omer. Behavioral and neurochemical alterations in rats prenatally exposed to 2,4-dichlorophenoxyacetate (2,4,d) and 2,4,5-trichlorophenoxyacetate (2,4,5-d) mixture. *Teratology*, 37(5):515–515, 1988.

[264] D. F. Molefe, J. J. Chen, P. C. Howard, B. J. Miller, C. P. Sambuco, P. D. Forbes, and R. L. Kodell. Tests for effects on tumor frequency and latency in multiple dosing photocarcinogenicity experiments. *Journal of Statistical Planning and Inference*, 129(1-2):39–58, 2005.

[265] G. Molenberghs and H. Geys. Multivariate clustered data analysis in developmental toxicity studies. *Statistica Neerlandica*, 55(3):319–345, 2001.

[266] H. Moon, H. Ahn, and R. L. Kodell. An age-adjusted bootstrap-based Poly-k test. *Statistics in Medicine*, 24(8):1233–1244, 2005.

[267] H. Moon, H. Ahn, and R. L. Kodell. A computational tool for testing dose-related trend using an age-adjusted bootstrap-based Poly-k test. *Journal of Statistical Software*, 16(7), 2006.

[268] D. F. Moore and A. Tsiatis. Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics*, 47(2):383–401, 1991.

[269] R. Morris. Developments of a water-maze procedure for studying spatial-learning in the rat. *Journal of Neuroscience Methods*, 11(1):47–60, 1984.

[270] D. Morton, P. N. Lee, J. S. Fry, W. R. Fairweather, J. K. Haseman, R. L. Kodell, J. J. Chen, A. J. Roth, and K. Soper. Statistical methods for carcinogenicity studies. *Toxicologic Pathology*, 30(3):403–414, 2002.

[271] U. Munzel and L. A. Hothorn. A unified approach to simultaneous rank test procedures in the unbalanced one-way layout. *Biometrical Journal*, 43(5):553–569, 2001.

[272] F.J. Murray, F.M. Sullivan, A.K. Tiwary, and S. Carey. 90-day subchronic toxicity study of sodium molybdate dihydrate in rats. *Regulatory Toxicology and Pharmacology*, 70(3):579–588, 2014.

[273] J. A. Murrell, C. J. Portier, and R. W. Morris. Characterizing dose-response I: Critical assessment of the benchmark dose concept. *Risk Analysis*, 18(1):13–26, 1998.

[274] J. S. Najita, Y. Li, and P. J. Catalano. A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 58:555–573, 2009.

[275] N. Nakanishi, T. Hashimoto, and C. Hamada. Consideration of robustness and power of the Williams multiple comparison test in the evaluation of practical pharmacology. *Journal of Pharmacological Sciences*, 121:68P–68P, 2013.

[276] M. Nazarov and H. Geys. New r routines for facilitating comet assay studies in toxicology m. In *Nonclinical Statistics Conference Brugge*, 2014.

[277] K. Neubert and E. Brunner. A studentized permutation test for the non-parametric Behrens-Fisher problem. *Computational Statistics and Data Analysis*, 51(10):5192–5204, 2007.

[278] M. Neuhauser, H. Buning, and L. Hothorn. Maximum test versus adaptive tests for the two-sample location problem. *Journal of Applied Statistics*, 31(2):215–227, 2004.

[279] M. C. Newman. "What exactly are you inferring?" A closer look at hypothesis testing. *Environmental Toxicology and Chemistry*, 27(5):1013–1019, 2008.

[280] H. Nishiyama, T. Omori, and I. Yoshimura. A composite statistical procedure for evaluating genotoxicity using cell transformation assay data. *Environmetrics*, 14(2):183–192, 2003.

[281] A.M. Nyman, K. Schirmer, and R. Ashauer. Toxicokinetic-toxicodynamic modelling of survival of Gammarus pulex in multiple pulse exposures to propiconazole: model assumptions, calibration data requirements and predictive power. *Ecotoxicology*, 21(7):1828–1840, 2012.

[282] OECD. Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. Technical report, OECD: Organization for Economic Cooperation and Development, Paris, France, pp 62–102, 2006.

[283] OECD407. Repeated dose 28-day oral toxicity study in rodents, updated guideline, adopted 3rd october 2008. Technical report, OECD Paris, 2008.

[284] OECD408. Repeated dose 90-day oral toxicity study in rodents,updated guideline, adopted 21st september 1998. Technical report, OECD Paris, 1998.

[285] OECD426. OECD guideline for the testing of chemicals. developmental neurotoxicity study. Technical report, OECD, 2007.

[286] OECD429. OECD guideline for testing of chemicals: Skin sensitisation: Local lymph node assay. Technical report, OECD/OCDE 429, 2002.

[287] OECD471. OECD guideline for testing of chemicals: Bacterial reverse mutation test. Technical report, OECD/OCDE 471, 1997.

[288] OECD474. OECD guideline for testing of chemicals: In vivo micronucleus test. Technical report, OECD/OCDE 474, 2006.

[289] OECD486. Unscheduled DNA synthesis (uds) test with mammalian liver cells in vivo. Technical report, OECD, 1997.

[290] OECD487. OECD guideline for testing of chemicals: In vitro micronucleus test. Technical report, OECD/OCDE 487, 2006.

[291] J. Ogawa. On the confidence-bounds of the ratio of the means of a bivariate normal-distribution. *Annals of the Institute of Statistical Mathematics*, 35(1):41–48, 1983.

[292] J. G. Orelien, J. Zhai, R. Morris, and R. Cohn. An approach to performing multiple comparisons with a control in GEE models. *Communications in Statistics-Theory and Methods*, 31(1):87–105, 2002.

[293] P. Pallmann. *toxbox: Boxplots for Toxicological Data*, 2015. R package version 1.1.3.

[294] P. Pallmann and L.A. Hothorn. Boxplots for grouped and clustered data in toxicology. *Archives of Toxicology*, DOI 10.1007/s00204-015-1608-4, 2015.

[295] P. Pallmann, M. Pretorious, and C. Ritz. Simultaneous comparisons of treatments at multiple time points: Combined marginal models versus joint modeling. *Statistical Methods in Medical Research*, 2015.

[296] C. Parfett, A. Williams, J. L. Zheng, and G. Zhou. Gene batteries and synexpression groups applied in a multivariate statistical approach to dose-response analysis of toxicogenomic data. *Regulatory Toxicology and Pharmacology*, 67(1):63–74, 2013.

[297] S. Paul and K. K. Saha. The generalized linear model and extensions: A review and some biological and environmental applications. *Environmetrics*, 18(4):421–443, 2007.

[298] S. R. Paul. Analysis of proportions of affected fetuses in teratological experiments. *Biometrics*, 38(2):361–370, 1982.

[299] S. D. Peddada and G. E. Kissling. A survival-adjusted quantal-response test for analysis of tumor incidence rates in animal carcinogenicity studies. *Environmental Health Perspectives*, 114(4):537–541, 2006.

[300] R. Peto, M.C. Pike, and N.E. . Day. Guidelines for simple sensitive significance tests for carcinogenic effects in long-term animal experiments. Technical report, IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Supplement 2: Long-term and Short-term Screening Assays for Carcinogens: A Critical Appraisal. Lyon: International Agency for Research on Cancer. 311-346., 1980.

[301] A. Phillips, C. Fletcher, G. Atkinson, E. Channon, A. Douiri, T. Jaki, J. Maca, D. Morgan, J. H. Roger, and P. Terrill. Multiplicity: Discussion points from the Statisticians in the Pharmaceutical Industry multiplicity expert group. *Pharmaceutical Statistics*, 12(5):255–259, 2013.

[302] W. W. Piegorsch. Translational benchmark risk analysis. *Journal of Risk Research*, 13(5):653–667, 2010.

[303] W. W. Piegorsch, L. L. An, A. A. Wickens, R. W. West, E. A. Pena, and W. S. Wu. Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmetrics*, 24(3):143–157, 2013.

[304] H. P. Piepho. An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics*, 13(2):456–466, 2004.

[305] H. P. Piepho, A. Buchse, and K. Emrich. A hitchhiker's guide to mixed models for randomized experiments. *Journal of Agronomy and Crop Science*, 189(5):310–322, 2003.

[306] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. *nlme: Linear and Nonlinear Mixed Effects Models*, 2014. R package version 3.1-117.

[307] C. B. Pipper, C. Ritz, and H. Bisgaard. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society Series C- Applied Statistics*, 61:315–326, 2012.

[308] J. P. H. T. M. Ploemen, H. Kramer, E. I. Krajnc, and I. Martin. The use of toxicokinetic data in preclinical safety assessment: A toxicologic pathologist perspective. *Toxicological Pathology*, 35(6):834–837, 2007.

[309] M. J. Podgor, J. L. Gastwirth, and C. R. Mehta. Efficiency robust tests of independence in contingency tables with ordered classifications. *Statistics in Medicine*, 15(19):2095–2105, 1996.

[310] C. J. Portier and A. J. Bailer. Testing for increased carcinogenicity using a survival-adjusted quantal response test. *Fundamental and Applied Toxicology*, 12(4):731–737, 1989.

[311] S. Pramana, D. Lin, P. Haldermans, Z. Shkedy, T. Verbeke, H. Gohlmann, A. De Bondt, W. Talloen, and L. Bijnens. Isogene: An r package for analyzing dose-response studies in microarray experiments. *R Journal*, 2(1):5–12, 2010.

[312] C. J. Price, C. A. Kimmel, J. D. George, and M. C. MARR. The developmental toxicity of diethylene glycol dimethyl ether in mice. *Fundamental and Applied Toxicology*, 8(1):115–126, 1987.

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

[314] National Toxicology Program. National toxicology program. Toxicology and carcinogenesis studies of methyleugenol in F344/n rats and B6C3F1 mice. Technical report, 2000, Technical Report 491.

[315] J. Ranke. *drfit: Dose-response data evaluation*, 2014. R package version 0.6.3.

[316] M. Razzaghi. A hierarchical model for the skew-normal distribution with application in developmental neurotoxicology. *Communications in Statistics-Theory and Methods*, 43(8):1859–1872, 2014.

[317] J. Reiczigel, D. Zakarias, and L. Rozsa. A bootstrap test of stochastic equality of two populations. *American Statistician*, 59(2):156–161, 2005.

[318] G. Reifferscheid, H. M. Maes, B. Allner, J. Badurova, S. Belkin, K. Bluhm, F. Brauer, J. Bressling, S. Domeneghetti, T. Elad, S. Flueckiger-Isler, H. S. Grummt, R. Guertler, A. Hecht, M. B. Heringa, H. Hollert, S. Huber, M. Kramer, A. Magdeburg, H. T. Ratte, R. Sauerborn-Klobucar, A. Sokolowski, P. Soldan, T. Smital, D. Stalter, P. Venier, C. Ziemann, J. Zipperle, and S. Buchinger. International round-robin study on the Ames fluctuation test. *Environmental and Molecular Mutagenesis*, 53(3):185–197, 2012.

[319] M. Rhodes, S. Laffan, C. Genell, J. Gower, C. Maier, T. Fukushima, G. Nichols, and A. E. Bassiri. Assessing a theoretical risk of dolutegravir-induced developmental immunotoxicity in juvenile rats. *Toxicological Sciences*, 130(1):70–81, 2012.

[320] C. Rigaud, C. M. Couillard, J. Pellerin, B. Legare, and P. V. Hodson. Applicability of the TCDD-TEQ approach to predict sublethal embryo-toxicity in Fundulus heteroclitus. *Aquatic Toxicology*, 149:133–144, 2014.

[321] R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C- Applied Statistics*, 54:507–544, 2005.

[322] C. Ritz. Benchmark Dose Analysis in R. Under preparation, 2009.

[323] C. Ritz, D. Gerhard, and L. A. Hothorn. A unified framework for benchmark dose estimation applied to mixed models and model averaging. *Statistics in Biopharmaceutical Research*, 5(1):79–90, 2013.

[324] C. Ritz and J. C. Streibig. Bioassay Analysis using R. *Journal of Statistical Software*, 12, 2005.

[325] C. Ritz and L. Van der Vliet. Handling nonnormality and variance heterogeneity for quantitative sublethal toxicity tests. *Environmental Toxicology and Chemistry*, 28(9):2009–2017, 2009.

[326] M. Royer, P. N. Diouf, and T. Stevanovic. Polyphenol contents and radical scavenging capacities of red maple (acer rubrum l.) extracts. *Food and Chemical Toxicology*, 49(9):2180–2188, 2011.

[327] G. Rucker, G. Schwarzer, J. Carpenter, and I. Olkin. Why add anything to nothing? the arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, 28(5):721–738, 2009.

[328] P. E. Rudolph. Robustness of multiple comparison procedures - treatment versus control. *Biometrical Journal*, 30(1):41–45, 1988.

[329] G. D. Ruxton. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, 17(4):688–690, 2006.

[330] E. Ryu. Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. *Statistics in Medicine*, 28(25):3179–3188, 2009.

[331] E. J. Ryu and A. Agresti. Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, 27(10):1703–1717, 2008.

[332] ICH S3a. Note for guidance on toxicokinetics: A guidance for assessing systemic exposure in toxicology studies (cpmp/ich/384/95). Technical report, CPMP/ICH, 1995.

[333] K. K. Saha, R. Bilisoly, and D. M. Dziuda. Hybrid-based confidence intervals for the ratio of two treatment means in the over-dispersed Poisson data. *Journal of Applied Statistics*, 41(2):439–453, 2014.

[334] K.K. Saha. Inference concerning a common dispersion of several treatment groups in the analysis of over/underdispersed count data. *Biometrical Journal*, 56(3):441–460, 2014.

[335] S. Sand, A. F. Filipsson, and K. Victorin. Evaluation of the benchmark dose method for dichotomous data: Model dependence and model selection. *Regulatory Toxicology and Pharmacology*, 36(2):184–197, 2002.

[336] S. Sand, D. von Rosen, P. Eriksson, A. Fredriksson, H. Viberg, K. Victorin, and A. F. Filipsson. Dose-response modeling and benchmark calculations from spontaneous behavior data on mice neonatally exposed to 2,2 ',4,4 ',5-pentabromodiphenyl ether. *Toxicological Sciences*, 81(2):491–501, 2004.

[337] S. Sasabuchi. A multivariate one-sided test with composite hypothesis when the covariance matrix is completely unknown. *Memoirs of the Faculty of Science, Series A*, 42:37–46, 1988.

[338] F. Schaarschmidt. Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. *Computational Statistics and Data Analysis*, 58:265–275, 2013.

[339] F. Schaarschmidt. *mixADA: Normalization, mixture models and screening cutpoints for anti-drug-antibody reactions (based on contributions by Bettina Gruen and Thomas Jaki and Ludwig Hothorn)*, 2014. R package version 1.2.

[340] F Schaarschmidt. One-sided ratio-to-control tests - simulation results. Technical report, Leibniz University Hannover. Institute of Biostatistics, 2014.

[341] F. Schaarschmidt and D. Gerhard. *pairwiseCI: Confidence intervals for two sample comparisons*, 2013. R package version 0.1-22.

[342] F. Schaarschmidt, D. Gerhard, and M. Sill. *MCPAN: Multiple comparisons using normal approximation*, 2013. R package version 1.1-15.

[343] F. Schaarschmidt, M. Hofmann, T. Jaki, B. Gruen, and L. A. Hothorn. Statistical approaches for the determination of cut points in anti-drug antibody bioassays. *Journal of Immunological Methods*, 418:84–100, 2015.

[344] F. Schaarschmidt and L. A. Hothorn. Statistical methods and software for validation studies on new in vitro toxicity assays. *ATLA-Alternatives to Laboratory Animals*, 42(5):318–325, November 2014.

[345] F. Schaarschmidt and L.A. Hothorn. A note on prediction intervals for comparisons of a future group of historical control data. Technical report, Reports of the Institute of Biostatistics 01/2015. Leibniz University Hannover, 2015.

[346] F. Schaarschmidt, M. Sill, and L. A. Hothorn. Approximate Simultaneous Confidence Intervals for Multiple Contrasts of Binomial Proportions. *Biometrical Journal*, 50(5, SI):782–792, OCT 2008.

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

[348] F. Schaarschmidt and L. Vaas. Analysis of trials with complex treatment structure using multiple contrast tests. *Hortscience*, 44(1):188–195, 2009.

[349] Frank Schaarschmidt. *binMto: Asymptotic simultaneous confidence intervals for many-to-one comparisons of proportions*, 2013. R package version 0.0-6.

[350] M. Scholze and A. Kortenkamp. Statistical power considerations show the endocrine disruptor low-dose issue in a new light. *Environmental Health Perspectives*, 115:84–90, 2007.

[351] L. Scrucca. qcc: an R package for quality control charting and statistical process control. *R News*, 4/1:11–17, 2004.

[352] D. Seidel. *Trendtests für geordnete kategoriale Daten bei sehr kleinen Fallzahlen.* PhD thesis, University of Hannover, 1999.

[353] R. S. Sellers, D. Morton, B. Michael, N. Roome, J. K. Johnson, B. L. Yano, R. Perry, and K. Schafer. Society of toxicologic pathology position paper: Organ weight recommendations for toxicology studies. *Toxicologic Pathology*, 35(5):751–755, 2007.

[354] S. Senn. Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25(24):4334–4344, 2006.

[355] E. A. C. Shirley and P. Newham. The choice between analysis of variance and analysis of covariance with special reference to the analysis of organ weights in toxicology studies. *Statistics in Medicine*, 3(1):85–91, 1984.

[356] E.A.C. Shirley. A nonparametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. *Biometrics*, 33(2):386–389, 1977.

[357] C. H. Sim, F. F. Gan, and T. C. Chang. Outlier labeling with boxplot procedures. *Journal of the American Statistical Association*, 100(470):642–652, 2005.

[358] R. T. Smythe, D. Krewski, and D. Murdoch. The use of historical control information in modeling dose-response relationships in carcinogenesis. *Statistics & Probability Letters*, 4(2):87–93, 1986.

[359] E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, 50(5, SI):641–656, 2008.

[360] M. Sprengel. *Analyse kategorialer Daten mit speziellem Fokus auf simultane Konfidenzintervalle.* PhD thesis, MASTERARBEIT zur Erlangung des Grades eines M.Sc. der Gartenbauwissenschaften der naturwissenschaftlichen Fakultaet an der Leipniz Universitaet Hannover, 2011.

[361] D. M. Stasinopoulos and R. A. Rigby. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7), 2007.

[362] K. E. Stebbins, K. A. Johnson, T. K. Jeffries, J. M. Redmond, K. T. Haut, S. N. Shabrang, and W. T. Stott. Chronic toxicity and oncogenicity studies of ingested 1,3-dichloropropene in rats and mice. *Regulatory Toxicology And Pharmacology*, 32(1):1–13, 2000.

[363] O. Sverdlov and W.Kee. Wong. Novel statistical designs for phase I/II and phase II clinical trials with dose-finding objectives. *Therapeutic Innovation & Regulatory Science*, 48(5):601–612, 2014.

[364] A. Swain, J. Turton, and C. et al. Scudamore. Nephrotoxicity of hexachloro-1:3-butadiene in the male Hanover Wistar rat; correlation of minimal histopathological changes with biomarkers of renal injury. *Journal of Applied Toxicology*, 32(6):417–428, 2012.

[365] A. Szabo. *CorrBin: Nonparametrics with clustered binary and multinomial data*, 2013. R package version 1.4.

[366] A. C. Tamhane and L. A. Hothorn. A multiple comparison procedure for three- and four-armed controlled clinical trials by M. A. Proschan in Statistics in Medicine 1999; 18 : 787-798. *Statistics in Medicine*, 20(2):317–318, 2001.

[367] M. C. Tamhane and B. R. Logan. A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. *Biometrika*, 91(3):715–727, September 2004.

[368] M. L. Tang, H. K. T. Ng, J. H. Guo, W. Chan, and B. P. S. Chan. Exact Cochran-Armitage trend tests: comparisons under different models. *Journal of Statistical Computation and Simulation*, 76(10):847–859, October 2006.

[369] R. E. Tarone. Tests for trend in life table analysis. *Biometrika*, 62(3):679–682, 1975.

[370] R.E. Tarone. The use of historical control information in testing for a trend in Poisson means. *Biometrics*, 38(2):457–462, 1982.

[371] P. Tattar. *gpk: 100 Data Sets for Statistics Education*, 2013. R package version 1.0.

[372] G.B. A Teuns, H. Geys, S.M. A. Geuens, P. Stinissen, and T.F. Meert. Abuse liability assessment in preclinical drug development: Predictivity of a translational approach for abuse liability testing using methylphenidate in four standardized preclinical study models. *Journal of Pharmacological and Toxicological Methods*, 70(3):295–309, 2014.

[373] Terry Therneau. *survival: Survival analysis, including penalised likelihood.*, 2011. R package version 2.36-9, <http://CRAN.R-project.org/package=survival>.

[374] T.M. Therneau. *coxme: Mixed Effects Cox Models*, 2015. R package version 2.2-4.

[375] T.M. Therneau, P.M. Grambsch, and V.S. Pankratz. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175, 2003.

[376] F. D. Toledo, L. M. Perez, C. L. Basiglio, J. E. Ochoa, E. J. S. Pozzi, and M. G. Roma. The ca2+-calmodulin-ca2+/calmodulin-dependent protein kinase ii signaling pathway is involved in oxidative stress-induced mitochondrial permeability transition and apoptosis in isolated rat hepatocytes. *Archives of Toxicology*, 88(9):1695–1709, 2014.

[377] Y. L. Tong. On partitioning a set of normal populations by their locations with respect to a control. *Annals of Mathematical Statistics*, 40(4):1300–1324, 1969.

[378] J. F. Troendle. A bootstrap test of stochastic equality of two populations - comment. *American Statistician*, 59(3):279–279, 2005.

[379] K.T. Tsai. Robust Williams trend test. *Communications in Statistics -Theory and Methods*, 29(5-6):1327–1346, 2000.

[380] J. W. Tukey, J. L. Ciminera, and J. F. Heyse. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics*, 41(1):295–301, 1985.

[381] F. Uibel, A. Muehleisen, C. Kohle, M. Weimer, T. C. Stummann, S. Bremer, and M. Schwarz. Reproglo: A new stem cell-based reporter assay aimed to predict embryotoxic potential of drugs and chemicals. *Reproductive Toxicology*, 30(1):103–112, 2010.

[382] L. N. Vandenberg. Non-monotonic dose responses in studies of endocrine disrupting chemicals: Bisphenol a as a case study. *Dose-response*, 12(2):259–276, 2014.

[383] P. Vasseur and C. Lasne. OECD Detailed Review Paper (DRP) number 31 on "Cell Transformation Assays for Detection of Chemical Carcinogens": Main results and conclusions. *Mutation Research-genetic Toxicology and Environmental Mutagenesis*, 744(1):8–11, 2012.

[384] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[385] P. E. Verde, L. A. Geracitano, L. L. Amado, C. E. Rosa, A. Bianchini, and J. M. Monserrat. Application of public-domain statistical analysis software for evaluation and comparison of comet assay data. *Mutation Research-Genetic Toxicology and Environmental Mutagenesis*, 604(1-2):71–82, 2006.

[386] W. J. Waddell. History of dose response. *Journal of Toxicological Sciences*, 35(1):1–8, February 2010.

[387] T. Waldmann, E. Rempel, N. V. Balmer, A. Konig, R. Kolde, J. A. Gaspar, M. Henry, J. Hescheler, A. Sachinidis, J. Rahnenfuhrer, J. G. Hengstler, and M. Leist. Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. *Chemical Research in Toxicology*, 27(3):408–420, 2014.

[388] D. I. Warton and F. K. C. Hui. The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, 92(1):3–10, 2011.

[389] M. Weimer, X. Q. Jiang, O. Ponta, S. Stanzel, A. Freyberger, and A. Kopp-Schneider. The impact of data transformations on concentration-response modeling. *Toxicology Letters*, 213(2):292–298, 2012.

[390] B. West, K. Welch, and A. Galecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman Hall / CRC Press, first edition, 2006. ISBN 1584884800.

[391] B. T. West and A. T. Galecki. An overview of current software procedures for fitting linear mixed models. *American Statistician*, 65(4):274–282, 2011.

[392] P. Westfall. Proc multtest: Example 58.4 Fisher test with permutation resampling. Technical report, SAS Inc., 2014.

[393] P. H. Westfall and S. S. Young. P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, 84(407):780–786, 1989.

[394] H. I. Weston, M. E. Sobolewski, J. L. Allen, D. Weston, K. Conrad, S. Pelkowski, G. E. Watson, G. Zareba, and D. A. Cory-Slechta. Sex-dependent and non-monotonic enhancement and unmasking of methylmercury neurotoxicity by prenatal stress. *Neurotoxicology*, 41:123–140, 2014.

[395] G. C. White and R. E. Bennetts. Analysis of frequency count data using the negative binomial distribution. *Ecology*, 77(8):2549–2557, 1996.

[396] H. Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.

[397] H. Wickham and P. Hadley. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 2007.

[398] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[399] S. J. Wiklund and E. Agurell. Aspects of design and statistical analysis in the comet assay. *Mutagenesis*, 18(2):167–175, 2003.

[400] D.A. Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1):103–117, 1971.

[401] D.A. Williams. The comparison of several dose levels with a zero dose control. *Biometrics*, 28(2):519–531, 1972.

[402] J. B. Wilson. Priorities in statistics, the sensitive feet of elephants, and don't transform data. *Folia Geobotanica*, 42(2):161–167, 2007.

[403] M. J. Wolfsegger, G. Gutjahr, W. Engl, and T. Jaki. A hybrid method to estimate the minimum effective dose for monotone and non-monotone dose-response relationships. *Biometrics*, 70(1):103–109, 2014.

[404] M. J. Wolfsegger and T. Jaki. Assessing systemic drug exposure in repeated dose toxicity studies in the case of complete and incomplete sampling. *Biometrical Journal*, 51(6):1017–1029, 2009.

[405] M. J. Wolfsegger and T. Jaki. Non-compartmental estimation of pharmacokinetic parameters in serial sampling designs. *Journal of Pharmacokinetics and Pharmacodynamics*, 36(5):479–494, 2009.

[406] M. J. Wolfsegger, T. Jaki, B. Dietrich, J. A. Kunzler, and K. Barker. A note on statistical analysis of organ weights in non-clinical toxicological studies. *Toxicology and Applied Pharmacology*, 240(1):117–122, 2009.

[407] G. H. Woo, M. Shibutani, T. Ichiki, M. Hamamura, K. Y. Lee, K. Inoue, and M. Hirose. 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. *Archives Of Toxicology*, 81(2):77–88, 2007.

[408] B. L. Wu and A. R. de Leon. Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology. *Journal of Agricultural Biological and Environmental Statistics*, 19(1):39–56, 2014.

[409] Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC- The R Series, 2015.

[410] E. Yamamoto and T. Yanagimoto. Statistical-methods for the beta-binomial model in teratology. *Environmental Health Perspectives*, 102:25–31, 1994.

[411] T. Yanagawa and Y. Kikuchi. Statistical issues on the determination of the no-observed-adverse-effect levels in toxicology. *Environmetrics*, 12(4):319–325, 2001.

[412] Thomas W. Yee. The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32(10):1–34, 2010.

[413] Thomas W. Yee. *VGAM: Vector Generalized Linear and Additive Models*, 2014. R package version 0.9-5.

[414] M. Yokohira, K. Hosokawa, K. Yamakawa, N. Hashimoto, S. Suzuki, Y. Matsuda, K. Saoo, T. Kuno, and K. Imaida. A 90-day toxicity study of l-asparagine, a food additive, in f344 rats. *Food and Chemical Toxicology*, 46(7):2568–2572, 2008.

[415] A. Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16, 2006.

[416] A. Zeileis, C. Kleiber, and Simon Jackman. Regression models for count data in R. *Journal of Statistical Software*, 27(8):8, 2008.

[417] G. Zheng. Analysis of ordered categorical data: Two score-independent approaches. *Biometrics*, 64(4):1276–1279, 2008.

[418] D. W. Zimmerman. Type I error probabilities df the Wilcoxon-Mann-Whitney test and student t test altered by heterogeneous variances and equal sample sizes. *Perceptual And Motor Skills*, 88(2):556–558, 1999.

[419] D. W. Zimmerman. A note on preliminary tests of equality of variances. *British Journal of Mathematical & Statistical Psychology*, 57:173–181, 2004.

[420] D. W. Zimmerman and B. D. Zumbo. Rank transformations and the power of the student t-test and welch t-test for nonnormal populations with unequal variances. *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale*, 47(3):523–539, 1993.

[421] A. Zuur, E.N. Ieno, N. Walker, A.A. Saveliev, and G.M. Smith. *Mixed Effects Models and Extensions in Ecology with R (Statistics for Biology and Health)*. Springer, 2009 edition, 3 2009.