# On the behavior of multiple comparison procedures in complex parametric designs

**Esther Herberich**


München 2012

# On the behavior of multiple comparison procedures in complex parametric designs

**Esther Herberich**

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Esther Herberich
aus Würzburg

München, den 28.08.2012

Erstgutachter: Prof. Dr. Torsten Hothorn
Zweitgutachter: Prof. Dr. Frank Bretz
Tag der mündlichen Prüfung: 31.10.2012

# Danksagung

Danke sagen möchte ich ...

... Torsten Hothorn für die engagierte Betreuung und fachliche Unterstützung in der Nähe und aus der Ferne.

... Frank Bretz für seine Bereitschaft diese Arbeit zu begutachten.

... Ludwig A. Hothorn für fachliche Anregungen und tiefere Einblicke in die Welt der simultanen Inferenz.

... Johannes Sikorski, Markus Pfirrmann und Christine Hassler für die Zusammenarbeit an den Anwendungsbeispielen.

... Nikolay Robinzonov für die Begleitung durch den Büroalltag über die letzten Jahre.

... meinen Kollegen vom Institut für Statistik für Unterhaltung und Ablenkung bei gemeinsamen Mittagessen, Spieleabenden und sportlichen Aktivitäten.

... Andreas, Martina und Verena für Anmerkungen und Verbesserungsvorschläge zu dieser Arbeit.

... Markus für die wunderbare Unterstützung auf den letzten Metern.

... meiner Familie insbesondere meinen Eltern für ihren Rückhalt und ihre Hilfe.

# Abstract

The framework for simultaneous inference by Hothorn, Bretz, and Westfall (2008) allows for a unified treatment of multiple comparisons in general parametric models where the study questions are specified as linear combinations of elemental model parameters. However, due to the asymptotic nature of the reference distribution the procedure controls the error rate across all comparisons only for sufficiently large samples. This thesis evaluates the small samples properties of simultaneous inference in complex parametric designs. These designs are necessary to address questions from applied research and include non-standard parametric models or data in which the assumptions of classical procedures for multiple comparisons are not met.

This thesis first treats multiple comparisons of samples with heterogeneous variances. Usage of a heteroscedastic consistent covariance estimation prevents an increase in the probability of false positive findings for reasonable sample sizes whereas the classical procedures show liberal or conservative behavior which persists even with increasing sample size.

The focus of the second part are multiple comparisons in survival models. Multiple comparisons to a control can be performed in correlated survival data modeled by a frailty Cox model under control of the familywise error rate in sample sizes applicable for clinical trials. As a further application, multiple comparisons in survival models can be performed to investigate trends. The procedure achieves good power to detect different dose-response shapes and controls the error probability to falsely detect any trend.

The third part addresses multiple comparisons in semiparametric mixed models. Simultaneous inference in the linear mixed model representation of these models yields an approach for multiple comparisons of curves of arbitrary shape. The sections on which curves differ can also be identified. For reasonably large samples the overall error rate to detect any non-existent difference is controlled. An extension allows for multiple comparisons of areas under the curve. However the resulting procedure achieves an overall error control only for sample sizes considerably larger than available in studies in which multiple AUC comparisons are usually performed.

The usage of the evaluated procedures is illustrated by examples from applied research including comparisons of fatty acid contents between Bacillus simplex lineages, comparisons of experimental drugs with a control for prolongation in survival of chronic myelogeneous leukemia patients, and comparisons of curves describing a morphological structure along the spinal cord between variants of the EphA4 gene in mice.

# Zusammenfassung

Das simultane Inferenzverfahren von Hothorn, Bretz und Westfall (2008) liefert einen einheitlichen Ansatz für Mehrfachvergleiche in allgemeinen parametrischen Modellen, in denen Fragestellungen anhand von Linearkombinationen der Modellparameter beschrieben werden. Da die Referenzverteilung, aus der der kritische Wert für die einzelnen Vergleiche berechnet wird, nur asymptotisch gilt, lässt sich die Kontrolle der Gesamtfehlerwahrscheinlichkeit über alle Vergleiche hinweg nur für hinreichend große Stichproben gewährleisten. In der vorliegenden Dissertation werden die Eigenschaften des simultanen Inferenzverfahren in komplexen parametrischen Designs und für kleine Stichproben untersucht. Multiple Vergleiche in komplexen parametrischen Designs werden in der angewandten Forschung benötigt, wenn die Datenstruktur durch einfache Modelle nicht ausreichend erfasst werden kann oder die von klassischen multiplen Testverfahren getroffenen Annahmen nicht erfüllt sind.

Im ersten Teil der Dissertation werden Mittelwertsvergleiche für Stichproben mit heterogenen Varianzen evaluiert. Mittels einer unter Heteroskedastizität konsistenten Kovarianzschätzung kann bereits mit moderaten Stichprobenumfängen die Wahrscheinlichkeit falsch positiver Ergebnisse gering gehalten werden. Klassische Verfahren für multiple Mittelwertsvergleiche hingegen zeigen liberales oder konservatives Verhalten, das auch mit zunehmendem Stichprobenumfang bestehen bleibt.

Im zweiten Teil der Dissertation werden multiple Vergleiche in Überlebenszeitmodellen betrachtet. Korrelierte Überlebenzeitdaten lassen sich mit einem Frailty Cox-Modell modellieren. Beim Vergleich der Überlebenszeiten zwischen mehreren experimentellen Therapien und einer Standardtherapie wird das vorgegebene $\alpha$-Niveau bei für Phase III Studien üblichen Stichprobenumfängen eingehalten. In Überlebenszeitmodellen finden multiple Vergleiche außerdem bei Trendanalysen Anwendung. Bei Verwendung einer geeigneten Kontrastmatrix können verschiedene Dosis-Wirkungs-Verläufe gut erkannt werden und nur mit der erlaubten Fehlerwahrscheinlichkeit $\alpha$ entscheidet das Verfahren für einen in Wirklichkeit nicht vorhandenen Trend.

Im dritten Teil werden gemischte semiparametrische Modelle betrachtet. Simultane Inferenz in der Darstellung dieser Modelle als lineare gemischte Modelle liefert einen Ansatz für multiple Vergleiche von Kurven beliebiger Formen. Der Ansatz liefert bei Vorliegen eines globalen Unterschieds zwischen den Kurven Aufschluss darüber, in welchen Bereichen sich die Kurven unterscheiden. Für hinreichend große Stichproben wird die Gesamtfehlerquote begrenzt. Eine Erweiterung des Verfahrens ermöglicht multiple Vergleiche von Flächen unter den Kurven. Dieses Verfahren kann jedoch nur dann die globale Fehlerquo-

te einhalten, wenn bedeutend größere Stichproben vorliegen, als dies üblicherweise in Studien, deren Fragestellungen sich mit multiplen AUC-Vergleichen beantworten lassen, der Fall ist.
Die Anwendung der untersuchten Verfahren wird an Beispielen aus der angewandten Forschung demonstriert. Darunter sind Vergleiche von *Bacillus simplex* Stämmen hinsichtlich ihres Fettsäuregehalts, Vergleiche von experimentellen Therapien bei chronischer myeloischer Leukämie mit der Standardtherapie und Vergleiche von Kurven, die eine morphologische Struktur entlang des Rückenmarks beschreiben, zwischen Mäusen mit verschiedenen Varianten des EphA4 Gens.

# Contents

# Chapter 1

# Introduction

Many research projects in the life sciences employ comparative studies, in which several samples are compared regarding a certain trait. Examples are the comparison of several treatments in efficacy studies in medical research, treatment of animals with different dosages of a compound in toxicity studies, comparisons of different mutations of a gene, or comparisons of several groups in repeated measures designs. Analyses of these studies aim to detect differences between groups or to identify an increasing or decreasing trend in dosage studies.
In comparisons of more than two groups not only the verification of a global difference or a global trend, but multiple pairwise comparisons are necessary in order to establish which groups differ or to detect the form of the dose-response relationship. When several groups are compared over time a further objective is to identify time points at which groups differ.
To draw valid conclusions across all comparisons made, a multiple Type I error level must be controlled. However, testing each single comparison on a significance level $\alpha$ increases the probability of a false positive on any of the tests above $\alpha$.
The overall Type I error rate can be held constant by adjusting the level of significance for each single test. Multiplicity adjustment procedures which lower the significance level for the single hypotheses are conservative, i.e. yield an overall error probability below the nominal level $\alpha$, if test statistics are correlated. Correlated test statistics arise from overlapping samples across pairwise contrasts, such as when multiple treatment groups are compared to the same control group. However, an overall Type I error rate close to $\alpha$ is desired, since conservative procedures detect existing differences with low probability or require high sample sizes to detect differences.

A selected class of multiple tests consists of so-called multiple contrast tests, which will be the focus of this thesis. Control of the overall error level across all contrasts, defined by linear combinations of model parameters, is achieved by incorporating the correlation among the test statistics into the critical values obtained from the joint distribution of the test statistics for all contrasts. The most common application of multiple contrast tests are comparisons of means between normally distributed samples with the all-pairwise com-

parison according to Tukey (1953) and the many-to-one comparison according to Dunnett (1955) as well-known examples. By adequate formulation of the multiple contrasts particular experimental questions can be adressed. Bretz, Genz, and Hothorn (2001) present multiple contrast tests and simultaneous confidence intervals for linear combinations of means in balanced and unbalanced designs under homoscedasticity. Munzel and Hothorn (2001) and Konietschke (2009) translate the multiple contrast tests to nonparametric settings.
Hothorn, Bretz, and Westfall (2008) extend the linear model framework of Bretz et al. (2001) to simultaneous inference procedures for parametric models with generally correlated parameter estimates. Applications include generalized linear models, linear and non-linear mixed effects models and survival models. The methods rely on asymptotic normality of the joint distribution of the test statistics for each contrast. However, due to the asymptotic nature of the reference distribution the procedure controls the multiple level across the family of comparisons only for sufficiently large samples.
The behavior of the simultaneous inference procedure by Hothorn et al. (2008) in small samples was evaluated for a variety of standard parametric models by Herberich (2009) in terms of the Type I error rate and power. Simulations indicate that the overall Type I error rate is generally well maintained even for moderate to small samples but that under certain scenarios the procedure is either conservative (e.g. for binary data) or liberal (e.g. for survival data).

The aim of this thesis is to evaluate the small samples properties of the simultaneous inference procedure by Hothorn et al. (2008) embedded in nonstandard parametric models or in designs in which the assumptions of classical procedures are not met. Scientific research provides several examples of investigations involving multiple comparisons in settings which cannot be adequately addressed by standard parametric models but require more complex designs. For relevant biological or medical research questions models are built which adequately model the available data structure and comparisons of multiple groups are performed within these models. To gauge whether valid conclusions can be drawn for these applications the small sample behavior of the procedure is investigated. The performance is measured by Type I error and Type II error rates generalized from the familiar error concepts for single null hypothesis to multiple tests. This thesis covers three applications:

In biological experiments the scientific hypothesis under test is often formulated in terms of mean differences among several groups. Frequently the variability differs between samples. The classical ANOVA model and standard post-hoc tests require normal data and homogeneous variances as a general assumption. If the homoscedasticity assumption is violated differences may be under- or overestimated, respectively, especially in unbalanced designs. For multiple comparisons of normal samples, several adjustments regarding heteroscedasticity have been proposed (e.g. Games and Howell, 1976; Brown and Forsythe, 1974a; Tamhane, 1977), but most of them only for special contrasts and/or showing conservative or liberal behavior depending on the kind of heteroscedasticity (Tamhane, 1979; Dunnett, 1980). Hasler and Hothorn (2008) generalize the multiple contrast tests by Bretz et al. (2001) to normal, heteroscedastic samples using Welch-type adjustment of

the degrees of freedom of the reference multivariate $t$-distribution. We propose to perform multiple comparisons of group means based on the framework by Hothorn et al. (2008) using a heteroscedastic-consistent sandwich estimator for the covariance matrix of parameter estimates from the general linear model. The performance of this approach is investigated in small normal and skewed samples.

The second field of multiple comparisons considered in this thesis are survival endpoints of several groups. This thesis adresses two questions common in applied research: In clinical trials many-to-one comparisons are of interest when survival times of the individuals under observation are not independent. In toxicological studies trends, and in particular trend shapes are one focus of research.
Clinical trials often compare the survival of patients undergoing several experimental therapies with that under one standard therapy. The traditional approach to model the effect of different treatments and other explanatory variables on patient survival is the Cox proportional hazards model. An increased number of false positive findings resulting from multiple comparisons is commonly prevented by the Bonferroni procedure. One of the central assumptions of the Cox proportional hazards model is that survival times of the individuals under observation are independent conditioned on the observed values of covariates. In some study designs, this assumption is not realistic, such as when study participants are recruited and treated in different study centers. Such correlated survival data can be modeled using frailty Cox models, in which random effects are added to the Cox proportional hazards model. On the basis of the framework for simultaneous inference by Hothorn et al. (2008) we consider many-to-one comparisons of treatments with adjustment for covariates for clustered survival data modeled by a frailty Cox model. The quality of the procedure is inspected in balanced and various unbalanced designs for samples sizes reasonable for phase III clinical trials.
A further research question addressed by multiple comparisons regarding survival endpoint arises in pharmacology, where toxicological studies compare survival (among other endpoints) of groups treated with increasing dosages of a chemical compound to survival of a control group to investigate a trend. The sample size is determined on the basis of regulatory guidelines and rather small due to ethical reasons. The lifetable test by Tarone (1975) is the approach recommended by the National Toxicology Program (NTP) of the U.S. for mortality trend analysis in long-term carcinogenicity bioassays. However, this procecdure only tests for deviations from a linear trend and is less sensitive to other dose-response-shapes. Tests on Williams-type contrasts test the equality of model parameters against an ordered alternative and are sensitive to several different dose-response shapes by successfully pooling the dosage groups and comparing the pooled samples to a control. Simultaneous inference for Williams-type contrasts is described for several endpoints such as normal and binomial data (Bretz and Hothorn, 2000, 2002). We consider simultaneous inference for Williams-type multiple contrasts when the response variable is time of survival employing the approach by Hothorn et al. (2008) in the Cox proportional hazards model or the accelerated failure time model with the objective of detecting a trend in mortality in toxicological studies. The performance of the procedure is evaluated in balanced and unbalanced designs with sample sizes according to the NTP guidelines.

The third part of this thesis adresses multiple curve comparisons. In many biological or medical experiments data arise as curves, such as growth curves, hormone level profiles, or antigen trajectories. Frequently these curves are to be compared across several groups. Current approaches for comparisons of two or more curves only allow to test for a global difference between two groups with multiplicity control over all pairs of groups compared (Zhang, Lin, and Sowers, 2000; Behseta and Chenouri, 2011; Kong and Yan, 2011). However, in some studies it is crucial to establish on which sections two curves differ if an overall difference is found. Pairwise comparisons of several curves over a grid along the curves result in a multiple testing problem, with the total number of tests equal to the number of pairwise comparisons of two groups multiplied with the number of positions on which the curves are to be compared. Semi-parametric mixed models provide a tool to describe smooth curves of unknown form using penalized splines with subject-specific deviations from the group-level curve modeled by random effects. Adapting the simultaneous inference procedure for general parametric models (Hothorn et al., 2008) to semi-parametric mixed models yields an approach for multiple curve comparisons providing information on the positions in which the curves differ. The behavior of multiple curve comparisons is evaluated in various designs differing with regard to the number of subjects per group and the number of measurements per subject including settings with censored observations.
An extension of the procedure for multiple curve comparisons offers a method for multiple comparisons of areas under curves of unknown form. In pharmacokinetics the area under a plasma concentration versus time curve measures the body exposure to drug after administration of a dose of the drug. Absent a kinetic model of the plasma concentration curve the area under the curve (AUC) is commonly estimated by applying the trapezoidal rule on the means of measurements at each time point. Pairwise comparisons of the AUC between two groups are performed using Student's $t$-test which requires multiplicity adjustment when used for comparisons of the AUC of more than two groups. This thesis introduces an alternative approach for multiple comparisons of areas under curves of unknown form. Group-level curves are fitted using a semi-parametric mixed model and estimates of the areas thereunder are expressed as contrasts of model parameter estimates. The areas under several curves can be compared employing the simultaneous inference procedure for general parametric models (Hothorn et al., 2008) in semi-parametric mixed models. The performance of the method for multiple AUC comparisons is evaluated by simulations.

In summary, this thesis evaluates the small sample behavior of simultaneous inference in complex models arising from applied research. Simultaneous inference in the described designs is performed by means of adjusted $p$-values for multiple contrast tests or by simultaneous confidence intervals which measure the magnitude of the differences. If the study objective is merely identification of a significant difference, hypotheses are assessed using adjusted $p$-values. In the designs with survival endpoint, in which hazard ratios oder survival time ratios represent interpretable effect sizes measuring the effects between groups, simultaneous confidence intervals are used which can in addition to assessing statistical significance be interpretated regarding clinical importance or biological relevance of a significant difference.

The present thesis is organized as follows: Relevant concepts of simultaneous inference are described in Chapter 2. Multiple contrast tests and simultaneous confidence intervals for general parametric models are introduced followed by the description of multiple testing concepts of Type I and Type II error rates and approaches for their estimation by simulations. In Chapter 3 the behavior of simultaneous inference procedures for heteroscedastic data in unbalanced designs is investigated for normal and skewed data. Chapter 4 evaluates the performance of simultaneous inference in survival models. Multiple comparisons of curves and the areas thereunder are addressed in Chapter 5. A summary is given in Chapter 6.

All calculations in this thesis were conducted using the statistical software R (R Development Core Team, 2012). The simultaneous inference procedure is implemented in the package **multcomp** (Hothorn et al., 2008). Heteroscedastic-consistent covariance estimations are provided in the package **sandwich** (Zeileis, 2004). Survival models are provided in the packages **survival** and **coxme** (Therneau, 2012a,b). Semiparametric mixed models are implemented in the package **mgcv** (Wood, 2006a).

# Chapter 2

# Simultaneous Inference in Parametric Models

This chapter introduces general concepts of multiple testing and describes, following the outline in Hothorn et al. (2008), the framework for simultaneous inference in general parametric models, which will in this thesis be evaluated in complex parametric designs. A more comprehensive overview of multiple comparison procedures is available in Bretz, Hothorn, and Westfall (2010).

## 2.1 The Problem of Multiple Testing

If the evaluation of a scientific question involves a statistical test problem with one single null hypothesis, for example the comparison of the effects of two treatments, the hypothesis is commonly assessed by a statistical testing procedure which controls the probability of incorrectly rejecting the null hypothesis at a significance level $\alpha$. This implies that the test yields a true-negative decision with a probability of $1 - \alpha$. If two null hypotheses are tested each at $\alpha = 0.05$, for instance to compare the effects of two conditions for two different endpoints, the probability of at least one false-positive result is $1-0.95^2 = 0.0975$ under the independence assumption. In other words, if no beneficial effect of one treatment over the other exists for any of the two endpoints, the probability of declaring a difference with respect to at least one outcome is 0.0975 and substantially larger than the nominal level $\alpha = 0.05$. The probability of at least one Type I error approaches 1 with the number of hypotheses $k$ increasing:

$$1 - (1 - \alpha)^k \overset{k\to\infty}{\longrightarrow} 1, \quad \alpha \in (0, 1].$$

To draw valid conclusions across all comparisons made, a multiple Type I error level needs to be controlled.

A version of the single hypothesis Type I error rate translated to simultaneous inference

about several hypotheses is the *familywise error rate (FWER)*. It measures the probability of committing at least one Type I error among all hypotheses tested:

$$\text{FWER} = \mathbb{P}(\text{Reject at least one partial hypothesis incorrectly})$$

The simplest procedure ensuring a familywise error rate of maximal $\alpha$ for hypotheses test problems involving $k$ null hypotheses is the Bonferroni correction, which lowers the level for each test to $1/k$ times the desired multiple level with $k$ the number of tests performed. The Bonferroni method is an example of a *single-step procedure* characterized by the fact that the decision on rejection or non-rejection of one null hypothesis does not consider or affect the decision on any other null hypothesis.

If multiple comparisons are performed on groups with overlap in the compared samples, e.g. all-pairwise comparisons of several groups, the test statistics involved in the $k$ null hypotheses are correlated. In this case, the Type I error rate without multiplicity adjustment is increased from the nominal level $\alpha$ of each test, but less than the probability $1 - (1 - \alpha)^k$ derived for independent test statistics. The Bonferroni method and other procedures which do not take dependencies between test statistics into account are conservative, i.e. yield a FWER below $\alpha$, and their power to detect existing differences is low so that large samples are needed.

Multiple contrast tests, the subject of this thesis, are a class of simultaneous inference procedures, which accommodate correlations between test statistics to prevent an inflation of the FWER.

## 2.2 Multiple Contrasts

Multiple contrast tests provide the advantage that comparisons can be defined most suitable to address the experimental questions. Each particular hypothesis is specified as a contrast, i.e. a linear combination of the model parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$. The constant coefficients of the linear combination are summarized in a contrast vector

$$\boldsymbol{c} = (c_1, \ldots, c_p).$$

A certain comparison is reflected by the pattern of positive and negative entries of $\boldsymbol{c}$ with the elements of $\boldsymbol{c}$ summing up to 0.
For illustration purposes let $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_p)^\top$ be the vector of expectations of $p$ random variables. If we are interested in comparing the expectations of the 1th and the $p$th group $\mu_1$ and $\mu_p$ the adequate linear combination $\boldsymbol{c\beta}$ is described by the contrast vector $\boldsymbol{c} = (-1, 0, \ldots, 0, 1) \in \mathbb{R}^{1\times p}$ leading to the null hypothesis

$$\mathrm{H}^0 : \boldsymbol{c\beta} = a,$$

which is equivalent to

$$\mathrm{H}^0 : \mu_p - \mu_1 = a$$

for $a \in \mathbb{R}$ the value under test.
Each separate comparison is covered by its corresponding contrast vector $\boldsymbol{c}_j = (c_{j1}, \ldots, c_{jp})$, $j = 1, \ldots, k$, so that $k$ questions lead to partial null hypotheses

$$\mathrm{H}_j^0 : \boldsymbol{c}_j \boldsymbol{\beta} = a_j, \quad j = 1, \ldots, k.$$

The contrast vectors can be summarized in a contrast matrix

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \\ \vdots \\ \boldsymbol{c}_k \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kp} \end{pmatrix}.$$

If we consider comparisons of the expectation of the first group with the expectations of any other group as an example, the contrast matrix is

$$\boldsymbol{C} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

specifying the global hypothesis

$$\mathrm{H}^0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{a}, \quad \boldsymbol{a} = (a_1, \ldots a_k).$$

The rows of $\boldsymbol{C}\boldsymbol{\beta}$ define the partial hypotheses

$$\begin{aligned} \mathrm{H}_1^0 &: \mu_2 - \mu_1 = a_1, \\ \mathrm{H}_2^0 &: \mu_3 - \mu_1 = a_2, \\ &\vdots \\ \mathrm{H}_k^0 &: \mu_p - \mu_1 = a_k. \end{aligned}$$

Examples of contrasts in parametric models are comparisons of means in ANOVA models, difference of effects on the hazard rate in the Cox proportional hazards model (corresponding to log hazard ratios), difference of effects on the log odds in the logit model (corresponding to log odds ratios). However, multiple contrasts are not restricted to comparing differences of raw model parameters. Comparisons of weighted model parameters can be expressed by linear combinations with the weights specified in the contrast vector.

## 2.3 Union Intersection Tests

Union intersection tests can be used as a foundation for the construction of multiple comparison procedures. One class of union intersection tests are max-$t$ tests, which will

be used for inference on hypotheses specified by multiple contrasts in Section 2.4.
Assume that we aim to detect an overall difference among the means of several groups. If $\mathrm{H}_j^0$, $j = 1, \ldots, k$, denote partial hypotheses each comparing the means of two groups with all possible pairs of groups compared in the $k$ partial hypotheses, one could claim an overall difference, if at least one partial hypothesis is rejected.
In general, the global null hypothesis can be expressed as the intersection of a family of null hypotheses $\bigcap_{j=1}^k \mathrm{H}_j^0$. Let $\mathcal{C}_j$ denote the rejection region of a test for $\mathrm{H}_j^0$, i.e., the $j$th partial hypothesis is rejected, if the associated test statistics $T_j$ is element of the set $\mathcal{C}_j$. The global null hypothesis can be rejected, if $T_j \in \mathcal{C}_j$ for any $j$, which means that the overall rejection region is the union of the rejection regions of all partial hypotheses $\bigcup_{j=1}^k \mathcal{C}_j$. With $\mathrm{H}_j^1$ denoting the alternative for $\mathrm{H}_j^0$ the overall test problem can then be expressed as

$$\mathrm{H}^0 : \bigcap_{j=1}^k \mathrm{H}_j^0 \quad \text{versus} \quad \mathrm{H}^1 : \bigcup_{j=1}^k \mathrm{H}_j^1.$$

### Max-$t$ Tests

Max-$t$ tests are a class of union intersection tests. Let again $T_j$ denote the test statistics for hypothesis $\mathrm{H}_j^0$, $j = 1, \ldots, k$. We assume without loss of generality that large values of the test statistics favor the alternatives $\mathrm{H}_j^1$. If we take the maximum over the single test statistics

$$T_{\max} = \max\{T_1, \ldots, T_k\}$$

one approach to assess the global null hypothesis is to reject $\mathrm{H}^0$ if $T_{\max}$ exceeds a critical value $c$ chosen such that the Type I error rate is controlled. If small values of $T_j$ favor the alternatives, the minimum of the test statistics is taken instead. For two-sided test problems, the maximum of the absolute values of the single test statistics

$$T_{\max} = \max\{|T_1|, \ldots, |T_k|\}$$

is used. The critical value $c$ can be derived from the joint distribution of $T_1, \ldots, T_k$.

Note that the rejection of the global null hypothesis of a union intersection test does not provide information about which partial hypotheses are false. Critical values for each test statistic, associated adjusted $p$-values or simultaneous confidence intervals are ways to overcome this.

## 2.4 Simultaneous Inference About Multiple Contrasts

Hothorn et al. (2008) provide a unified description of test procedures for multiple contrasts in parametric models with generally correlated parameter estimates, which will be described in the following. In Subsection 2.4.1, the joint distribution of test statistics

for a set of multiple contrast hypotheses is derived. The calculation of the critical value and of adjusted $p$-values for each partial hypothesis, and the construction of simultaneous confidence intervals from the joint distribution of the test statistics is described in Subsections 2.4.2–2.4.4.

### 2.4.1 Asymptotic Distribution of Estimated Contrasts in General Parametric Models

Let

$$\mathcal{M}(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\eta})$$

denote a (semi-)parametric model with observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, an unknown vector of model parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ and a vector of random or nuisance parameters $\boldsymbol{\eta}$ if applicable. Our multiple comparisons of interest are contrasts $\boldsymbol{C\beta}$ specified by the contrast matrix $\boldsymbol{C} \in \mathbb{R}^{k \times p}$. Assume that an estimate $\hat{\boldsymbol{\beta}}_n \in \mathbb{R}^p$ for $\boldsymbol{\beta}$ can be obtained from the observations $(Y_1, \ldots, Y_n)$ with associated covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. If the assumptions of the central limit theorem are fulfilled $\hat{\boldsymbol{\beta}}_n$ is asymptotically multivariate normal

$$a_n^{1/2} \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right) \xrightarrow{d} N_p(0, \boldsymbol{\Sigma})$$

for some positive, nondecreasing sequence $a_n$.
Provided $\hat{\boldsymbol{\Sigma}}_n \in \mathbb{R}^{p \times p}$ is a consistent estimation of $\boldsymbol{\Sigma}$

$$a_n \hat{\boldsymbol{\Sigma}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}$$

the estimate $\hat{\boldsymbol{\beta}}_n$ is asymptotically multivariate normal

$$\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} N_p(\boldsymbol{\beta}, \hat{\boldsymbol{\Sigma}}_n).$$

According to Serfling (1980, Theorem 3.3) our contrasts of interest $\boldsymbol{C}\hat{\boldsymbol{\beta}}_n \in \mathbb{R}^k$ again is asymptotically multivariate normal

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_n \overset{a}{\sim} N_k(\boldsymbol{C\beta}, \underbrace{\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_n\boldsymbol{C}^T}_{=: \hat{\boldsymbol{\Sigma}}_n^*}) \quad \text{with}$$

$$a_n \hat{\boldsymbol{\Sigma}}_n^* \xrightarrow{\mathbb{P}} \underbrace{\boldsymbol{C\Sigma C}^T}_{=: \boldsymbol{\Sigma}^*}.$$

The diagonal elements of $\hat{\boldsymbol{\Sigma}}_n^*$ are assumed to be positive. The standardization of the contrasts of interest $\boldsymbol{C}\hat{\boldsymbol{\beta}}_n$, denoted by $\boldsymbol{T}_n$, again is asymptotically normal

$$\boldsymbol{T}_n = \boldsymbol{D}_n^{-1/2} \left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_n - \boldsymbol{C\beta}\right) \overset{a}{\sim} N_k(0, \boldsymbol{R}_n). \tag{2.1}$$

$\boldsymbol{D}_n = \operatorname{diag}(\hat{\boldsymbol{\Sigma}}_n^*)$ contains the variances of $\boldsymbol{C}\hat{\boldsymbol{\beta}}_n$ on the diagonal, and $\boldsymbol{R}_n$ denotes the correlation matrix of the standardized test statistics $\boldsymbol{T}_n$

$$\boldsymbol{R}_n = \boldsymbol{D}_n^{-1/2}\, \hat{\boldsymbol{\Sigma}}_n^*\, \boldsymbol{D}_n^{-1/2}.$$

Following Hothorn et al. (2008) the distribution of $\boldsymbol{T}_n$ converges to

$$\boldsymbol{T}_n = \boldsymbol{D}_n^{-1/2} \left( \boldsymbol{C}\hat{\boldsymbol{\beta}}_n - \boldsymbol{C}\boldsymbol{\beta} \right) = \left( a_n \, \boldsymbol{D}_n \right)^{-1/2} a_n^{1/2} \left( \boldsymbol{C}\hat{\boldsymbol{\beta}}_n - \boldsymbol{C}\boldsymbol{\beta} \right) \xrightarrow{d} N_k(0, \boldsymbol{R}).$$

The asymptotic distribution of $\boldsymbol{T}_n$ can be used for simultaneous inference on hypotheses described by $\boldsymbol{C}\boldsymbol{\beta}$, if convergence of multivariate probabilities calculated for the vector $\boldsymbol{T}_n$ is provided when $\boldsymbol{T}_n$ is assumed normally distributed with $\boldsymbol{R}_n$ treated as if it were the true correlation matrix. According to Hothorn et al. (2008) this condition holds since the probabilities $\mathbb{P}(\max(|\boldsymbol{T}_n| \leq l))$ are continuous functions in $\boldsymbol{R}_n$ and a critical value $l$ and $\boldsymbol{R}_n \xrightarrow{\mathbb{P}} \boldsymbol{R}$ as a consequence of Theorem 1.7 in Serfling (1980).

### 2.4.2 Critical Values for Multiple Contrast Tests

Following Section 2.4.1, the individual test statistics $\boldsymbol{T}_n = (T_1, \ldots, T_k)$ asymptotically follow

$$\boldsymbol{T}_n = \boldsymbol{D}_n^{-1/2} \left( \boldsymbol{C}\hat{\boldsymbol{\beta}}_n - \boldsymbol{C}\boldsymbol{\beta} \right) \overset{a}{\sim} N_k(0, \boldsymbol{R}_n) \tag{2.2}$$

under the condition of the global null hypothesis

$$\mathrm{H}^0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{a}.$$

The linear combinations $\boldsymbol{c}_j\boldsymbol{\beta}$, with $\boldsymbol{c}_j$ the rows of the contrast matrix $\boldsymbol{C}$, reflect the comparisons of interest. A critical value, common for all partial hypotheses, can be derived based on the asymptotic joint distribution of $\boldsymbol{T}_n$ such that the multiple level $\alpha$ is controlled.
The max-$t$ approach for a multiple testing procedure considers the maximum of individual test statistics $T_{\max}$, if large values of the test statistics favor the alternative. The critical value $l$ for rejection of the global hypothesis, and for rejection of the partial hypotheses is chosen such, that

$$\mathbb{P}(T_{\max} > l) = \alpha$$

or equivalently

$$\mathbb{P}(T_{\max} \leq l) = 1 - \alpha.$$

For one-sided test problems

$$\mathrm{H}_j^0 : \boldsymbol{c}_j\boldsymbol{\beta} \leq a_j, \quad j = 1, \ldots, k,$$

large values of the estimated standardized contrasts $T_1, \ldots, T_k$ favor the alternative. Thus, the critical value $l$ needs to be chosen such, that

$$\mathbb{P}(\max(\boldsymbol{T}_n) \leq l) \cong \int_{-\infty}^{l} \cdots \int_{-\infty}^{l} \varphi_k(x_1, \ldots, x_k; \boldsymbol{R}) \, dx_1 \cdots dx_k = 1 - \alpha,$$

where $\varphi_k$ denotes the density function of the $k$-dimensional distribution of $\boldsymbol{T}_n$. The resulting critical value is the lower one-sided $(1 - \alpha)$ quantile of the distribution of $\boldsymbol{T}_n$ and can be computed by numerical integration routines for multivariate normal distributions

described in Genz and Bretz (1999, 2002) and Bretz et al. (2001). The consistent estimation $\boldsymbol{R}_n$ can be plugged-in for the commonly unknown correlation matrix $\boldsymbol{R}$.
Let $\boldsymbol{t} = (t_1, \ldots, t_k)$ be the vector of observed test statistics. In the one-sided case a partial hypothesis $\mathrm{H}_j^0 : \boldsymbol{c}_j\boldsymbol{\beta} \leq a_j$ is rejected if

$$t_j > u_{1-\alpha}^l,$$

with $u_{1-\alpha}^l$ denoting the lower one-sided $(1-\alpha)$ quantile of the distribution of $\boldsymbol{T}_n$.

For two-sided test problems

$$\mathrm{H}_j^0 : \boldsymbol{c}_j\boldsymbol{\beta} = a_j, \quad j = 1, \ldots, k,$$

large values of the absolute values of the estimated standardized contrasts $|T_1|, \ldots, |T_k|$ favor the alternative. Thus, the critical value $l$ needs to be chosen such, that

$$\mathbb{P}(\max|\boldsymbol{T}_n| \leq l) \cong \int_{-l}^{l} \cdots \int_{-l}^{l} \varphi_k(x_1, \ldots, x_k; \boldsymbol{R})\, dx_1 \cdots dx_k = 1 - \alpha.$$

The resulting critical value is the two-sided $(1-\alpha)$ quantile of the distribution of $\boldsymbol{T}_n$.
Let again $\boldsymbol{t} = (t_1, \ldots, t_k)$ be the vector of observed test statistics. In the two-sided case a partial hypothesis $\mathrm{H}_j^0 : \boldsymbol{c}_j\boldsymbol{\beta} = a_j$ is rejected if

$$|t_j| > u_{1-\alpha}^{ts},$$

with $u_{1-\alpha}^{ts}$ denoting the two-sided $(1-\alpha)$ quantile of the distribution of $\boldsymbol{T}_n$.

According to the characteristics of union intersection tests the global null hypothesis $\mathrm{H}^0$ is rejected, if any partial hypothesis $\mathrm{H}_j^0$ is rejected. On the other hand, if $\mathrm{H}^0$ is rejected, at least one $\mathrm{H}_j^0$ must be rejected.

### 2.4.3 Adjusted $p$-Values

Adjusted $p$-values for each partial hypothesis incorporate the multiplicity adjustment and hence are directly comparable with the nominal significance level $\alpha$. For multiple contrast tests adjusted $p$-values achieve control of the multiple significance level $\alpha$ by using the multivariate normal distribution from equation (2.1) as reference distribution of the test statistics. Let $\boldsymbol{t} = (t_1, \ldots, t_k)$ be the vector of observed test statistics.
For one-sided partial hypotheses

$$\mathrm{H}_j^0 : \boldsymbol{c}_j\boldsymbol{\beta} \leq a_j$$

the associated adjusted $p$-value $p_j$ is given by

$$p_j = \mathbb{P}(\max(\boldsymbol{T}_n) > t_j) = 1 - \int_{-\infty}^{t_j} \cdots \int_{-\infty}^{t_j} \varphi_k(x_1, \ldots, x_k; \boldsymbol{R})\, dx_1 \cdots dx_k, \quad j = 1, \ldots, k.$$

For two-sided partial hypotheses

$$\mathrm{H}_j^0 : \boldsymbol{c}_j\boldsymbol{\beta} = a_j$$

the associated adjusted $p$-value $p_j$ is given by

$$p_j = \mathbb{P}(\max |\boldsymbol{T}_n| > |t_j|) = 1 - \int_{-|t_j|}^{|t_j|} \cdots \int_{-|t_j|}^{|t_j|} \varphi_k(x_1, \ldots, x_k; \boldsymbol{R})\, dx_1 \cdots dx_k, \quad j = 1, \ldots, k.$$

### 2.4.4 Simultaneous Confidence Intervals

Simultaneous $(1 - \alpha)$ confidence intervals have a joint coverage probability of at least $1 - \alpha$ for the multiple contrasts of interest.

Lower simultaneous $(1 - \alpha)$ confidence intervals specifying a lower bound for the multiple contrasts summarized in $\boldsymbol{C\beta}$ are given by

$$\left[\boldsymbol{C}\hat{\boldsymbol{\beta}}_n - u^l_{1-\alpha} \operatorname{diag}(\boldsymbol{D}_n)^{1/2}, \infty\right)$$

with $u^l_{1-\alpha}$ the lower one-sided $(1 - \alpha)$ quantile of the distribution of $\boldsymbol{T}_n$.

Two-sided simultaneous $(1 - \alpha)$ confidence intervals specifying a lower and an upper bound for the multiple contrasts summarized in $\boldsymbol{C\beta}$ are given by

$$\left[\boldsymbol{C}\hat{\boldsymbol{\beta}}_n - u^{ts}_{1-\alpha} \operatorname{diag}(\boldsymbol{D}_n)^{1/2}, \boldsymbol{C}\hat{\boldsymbol{\beta}}_n + u^{ts}_{1-\alpha} \operatorname{diag}(\boldsymbol{D}_n)^{1/2}\right]$$

with $u^{ts}_{1-\alpha}$ the two-sided $(1 - \alpha)$ quantile of the distribution of $\boldsymbol{T}_n$.

## 2.5 Selected Contrasts

The rich literature on multiple testing has yielded a variety of contrasts which are frequently used and which thus will be introduced briefly along with their corresponding effect differences resulting by multiplying the parameter estimates by the contrast matrix. The following contrasts are frequently employed in applied research and will be reconsidered in the following chapters. Let $\mathcal{M}(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\eta})$ be a general (semi-)parametric model. For sake of simplicity we present the contrast matrices when the parameter vector only contains the effects of a grouping variable with $M$ levels, i.e., $\boldsymbol{\beta} \in \mathbb{R}^M$. Further covariates, which are not of interest in the comparisons, would each be represented as an additional column vector of zeros in the contrast matrix.

### Dunnett

Dunnett-type contrasts are used to describe many-to-one comparisons, i.e., comparisons of several groups with a control group (Dunnett, 1955). An application is the comparison of two or more experimental treatments with a standard treatment in clinical trials. The associated contrast matrix is of the form

$$\boldsymbol{C}_{\text{Dunnett}} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ -1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

leading to effect differences

$$\boldsymbol{C}_{\text{Dunnett}}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_3 - \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_M - \boldsymbol{\beta}_1 \end{pmatrix} \widehat{=} \begin{matrix} \text{group } 2 & \text{vs.} & \text{group } 1 \\ \text{group } 3 & \text{vs.} & \text{group } 1 \\ \vdots & & \\ \text{group } M & \text{vs.} & \text{group } 1. \end{matrix}$$

## Tukey

Tukey-type contrasts are used for comparisons of all possible pairs of groups and usually tested two-sided (Tukey, 1953). The associated contrast matrix is of the form

$$\boldsymbol{C}_{\text{Tukey}} = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ & & & & \vdots & & & \\ -1 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}$$

leading to effect differences

$$\boldsymbol{C}_{\text{Tukey}}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_3 - \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_M - \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_M - \boldsymbol{\beta}_{M-1} \end{pmatrix} \widehat{=} \begin{matrix} \text{group } 2 & \text{vs.} & \text{group } 1 \\ \text{group } 3 & \text{vs.} & \text{group } 1 \\ & \vdots & \\ \text{group } M & \text{vs.} & \text{group } 1 \\ & \vdots & \\ \text{group } M & \text{vs.} & \text{group } M-1. \end{matrix}$$

## Williams

The trend test by Williams (1971, 1972) investigates monotone (either increasing or decreasing) trends with increasing dosages compared to a control and was formulated as an

approximate multiple contrast test by Bretz (1999). Each of the Williams-type contrasts corresponds to a certain dose-response shape. Starting from a comparison of the highest group with the control, in each step the next highest group is included in the analysis and the pooled average of the effects of the highest groups is compared to the effect of the control. This procedure leads to a higher power compared to Dunnett-type comparisons since the sample size increases. The associated contrast matrix is of the form

$$\boldsymbol{C}_{\text{Williams}} = \begin{pmatrix} -1 & 0 & \cdots & 0 & 0 & 1 \\ -1 & 0 & \cdots & 0 & \dfrac{n_{I-1}}{n_{I-1}+n_I} & \dfrac{n_I}{n_{I-1}+n_I} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ -1 & \dfrac{n_2}{\sum_{m=2}^{M} n_m} & \cdots & \dfrac{n_{I-2}}{\sum_{m=2}^{M} n_m} & \dfrac{n_{I-1}}{\sum_{m=2}^{M} n_m} & \dfrac{n_I}{\sum_{m=2}^{M} n_m} \end{pmatrix}$$

leading to effect differences

$$\boldsymbol{C}_{\text{Williams}}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_M - \boldsymbol{\beta}_1 \\ \dfrac{n_I}{n_{I-1}+n_I}\boldsymbol{\beta}_M + \dfrac{n_{I-1}}{n_{I-1}+n_I}\boldsymbol{\beta}_{M-1} - \boldsymbol{\beta}_1 \\ \vdots \\ \dfrac{n_I}{\sum_{m=2}^{M} n_m}\boldsymbol{\beta}_M + \dfrac{n_{I-1}}{\sum_{m=2}^{M} n_m}\boldsymbol{\beta}_{M-1} + \dfrac{n_{I-2}}{\sum_{m=2}^{M} n_m}\boldsymbol{\beta}_{M-2} + \ldots + \dfrac{n_2}{\sum_{m=2}^{M} n_m}\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \end{pmatrix} \widehat{=}$$

$$\begin{array}{lll} \text{group } M & \text{vs.} & \text{group 1} \\ \text{groups } M,\, M-1 & \text{vs.} & \text{group 1} \\ & \vdots & \\ \text{groups } M,\, M-1,\, M-2,\, \ldots,\, 2 & \text{vs.} & \text{group 1.} \end{array}$$

## 2.6 Estimation of Multiple Comparison Procedure Properties

The simultaneous inference procedure for multiple contrasts in general parametric models assumes the availability of an asymptotic normal parameter estimate and a consistent estimation of the associated covariance matrix. Provided these assumptions are met, the simultaneous inference procedure based on the asymptotic normality of the parameter estimates achieves a multiple level of $\alpha$ for sample sizes approaching infinity. However, in samples of finite size $n$ the actual multiple level of $\alpha$ can deviate considerably from $\alpha$ when the distribution of the test statistics differs from the asymptotic distribution (2.2). Such deviations can be caused by bias of parameter estimates, bias of associated covariance estimates or deviations from the asymptotic multivariate normality.
To gauge the influence such biases have on multiple comparisons in small samples by simulation studies requires the a priori definition of criteria for performance. In the following, selected error measures of multiple comparison procedures are defined and their estimation by simulations to investigate the properties of multiple comparison procedures in samples of finite size is described. An overview of the following and further error concepts is given in Westfall, Tobias, Rom, Wolfinger, and Hochberg (1999).

### Type I Error

The *familywise error rate* defined in Section 2.1 is the probability of any false positive finding among all hypotheses tested and used as measure of the multiple level $\alpha$. Strong control of the familywise error rate implies control of the Type I error rate under any configuration of false and true null hypotheses, i.e., the global null hypothesis is not required to be true (Hochberg and Tamhane, 1987). However, in the following chapters data sets are simulated under the global null hypothesis for estimation of the FWER. The FWER is estimated by the portion of datasets, in which at least one adjusted $p$-value associated with a true null hypothesis is below the nominal value of $\alpha$.
Transferring the coverage probability of a single confidence interval to interval-based simultaneous inference leads to the *joint coverage probability*

$$\mathbb{P}(\text{Each simultaneous confidence interval covers the corresponding true effect difference})$$

as measure of the multiple confidence level $1 - \alpha$. The joint coverage probability can be estimated via simulations by the portion of data sets in which the simultaneous confidence intervals cover the contrasts under which the data sets were simulated.

### Type II Error

For a simultaneous inference procedure with control of the FWER minimization of the Type II error and hence maximization of the power is required. Several generalizations of the single hypothesis power definition exist to measure the ability of a multiple testing

procedure to detect false hypotheses. The appropriate power definition needs to be chosen depending on the objective of each particular study. The two concepts of power which are used in this thesis are *individual power* and *disjunctive power*.
The individual power is defined as

$$\mathbb{P}(\text{Reject a particular false } H_j^0)$$

and measured for each false hypothesis (Westfall et al., 1999).
The disjunctive power is defined as

$$\mathbb{P}(\text{Reject at least one false } H_j^0)$$

and should be used when it is of interest to detect at least one effect (Senn and Bretz, 2007). When the effects of the alternative tend towards the effects under the null, the disjunctive power tends towards the FWER.

In the following, the estimated FWER is used as measure of the multiple Type I error rate of multiple contrast tests. The joint coverage probability is used as measure of the simultaneous confidence level of simultaneous confidence intervals for multiple contrasts. The definition of power used depends on the study objective.

# Chapter 3

# Multiple Comparisons of Group Means Under Heteroscedasticity

Standard approaches for multiple comparisons of group means assume homogeneous variances as a general rule. This condition is mathematically convenient, but rather unrealistic in applied research, where heteroscedasticity is likely to exist among samples from different environments or experimental conditions. In this chapter based on Herberich, Sikorski, and Hothorn (2010), the behavior of simultaneous inference according to Hothorn et al. (2008) is investigated for multiple comparisons of means from heteroscedastic samples. Robust sandwich estimators, which are consistent under heteroscedasticity, are employed for covariance estimation in the linear model. The performance of the approach is investigated for normal and skewed data in unbalanced designs.

Many research projects in the life sciences employ comparative studies. For example, biodiversity exploration in population genetics measures the properties of individuals belonging to different groups. Frequently, multiple groups are compared on traits which may differ quantitatively. The scientific hypothesis under test is then most often formulated in terms of mean differences among at least two of these groups. The researcher often cannot assume that variances are equal under all experimental conditions. Standard parametric procedures for comparisons of means, such as the methods by Tukey (1953) and Dunnett (1955), assume homogeneous variances among all groups. Applying these methods under heteroscedasticity, which refers to heterogeneous or unequal variances among the groups, can result in a probability for false positive results far higher than $\alpha$, especially when unequal group sizes and/or non-normally distributed data are present. Unfortunately, unequal variances, skewed data and unbalanced group sizes are realistic and hardly avoidable situations in applied research. A switch to non-parametric tests is not necessarily an option because even though they do not assume normality, they still assume that the shapes of the distributions are the same in all groups, which implies that variances are equal (Hollander and Wolfe, 1999). Several approaches for global comparison of several means under heteroscedasticity based on $t$ distributions with adjusted

degrees of freedom (Satterthwaite, 1946) or based on $F$-distributions (Welch, 1951; Brown and Forsythe, 1974b; Weerahandi, 1995; Lee and Ahn, 2003; Xu and Wang, 2008) have been proposed. Methods for multiple pairwise comparisons in presence of heteroscedasticity are described mainly for selected types of contrasts (Brown and Forsythe, 1974a; Tamhane, 1977; Games and Howell, 1976), most of them with conservative or liberal behavior depending on the extent of heteroscedasticity (Tamhane, 1979; Dunnett, 1980). Hasler and Hothorn (2008) extended the multiple contrast procedures for normally distributed samples with homogeneous variances (Bretz et al., 2001) to the heteroscedastic case using multivariate $t$-distributions with comparison-specific degrees of freedom and a correlation matrix depending on the sample variances as reference distribution.

The framework for simultaneous inference by Hothorn et al. (2008) can be applied for multiple comparisons in ANOVA models. If the parameter estimates are asymptotically normal and a consistent estimation of the associated covariance matrix is provided, multiple comparisons in balanced and unbalanced models with arbitrary error distribution and hence arbitrary data distribution and variance structure can be performed.

In the presence of heteroscedasticity the ordinary least squares (OLS) parameter estimates remain unbiased and asymptotically normal (Eicker, 1963). The usual estimator of the associated covariance matrix, however, is not consistent under unequal error variances. Several covariance estimations that are consistent under both homoscedasticity and heteroscedasticity of unknown form have been proposed (White, 1980; MacKinnon and White, 1985; Cribato-Neto, 2004; Cribato-Neto, Souza, and Vasconcellos, 2007; Cribato-Neto and da Silva, 2011).

In the following, the properties of the simultaneous inference procedure by Hothorn et al. (2008) using a heteroscedastic-consistent sandwich matrix as covariance estimator of the OLS parameter estimates are investigated for multiple comparisons under heteroscedasticity. Designs under homoscedasticity as well as under heteroscedasticity for normal and for skewed data are investigated in simulations. Familywise error rate and power are estimated for small sample sizes using different heteroscedastic-consistent covariance estimations and compared to the properties of multiple comparison procedures assuming homoscedasticity. We then reanalyze data from biodiversity research, where the multiple cladogenic splits of evolutionary lineages (putative ecotypes) of the bacterium *Bacillus simplex* as an adaptive response to the microclimatically heterogeneous environment of "Evolution Canyon", Israel, are being studied (Sikorski and Nevo, 2005; Sikorski, Pukall, and Stackebrandt, 2008b; Koeppel, Perry, Sikorski, Krizanc, Warner, and Ward, 2008; Sikorski, Brambilla, Kroppenstedt, and Tindall, 2008a). In this model population, unbalanced groups with frequently heterogeneous variances in their phenotypic properties are found. We conduct multiple comparisons and account for the existing heteroscedasticity. The analyses are additionally conducted with methods requiring homogeneous variances. For several comparisons the results differ depending on whether heterogeneous variances are accounted for. When neglecting the heteroscedasticity, in several comparisons significant differences are found although they are actually not present or significant differences are not detected although they are present when the appropriate method is used.

# 3.1 Simultaneous Inference Under Heteroscedasticity

## 3.1.1 Model and Hypotheses

We consider a one-way, fixed effects analysis of variance model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, M,\ j = 1, \ldots, n_i, \tag{3.1}$$

where $y_{ij}$ denotes the $j$th observation in group $i$, $\mu$ is the overall mean, $\alpha_i$ denotes the main effect in group $i$ with constraint $\sum_{i=1}^{M} \alpha_i = 0$, and $\varepsilon_{ij}$ are independent random errors with $\mathrm{Var}(\varepsilon_{ij}) = \sigma_i^2$.

To assess which particular groups differ in their means, we consider all pairwise comparisons of group effects

$$\mathrm{H}_{i\imath}^0 : \alpha_i - \alpha_\imath = 0 \quad \forall i \neq \imath,\ i, \imath = 1, \ldots, M. \tag{3.2}$$

Model (3.1) can be written as a general linear model

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \varepsilon_{ij} \tag{3.3}$$

with $\boldsymbol{x}_{ij}$ a $M$-dimensional vector with entry 1 at the $i$th position and entries 0 at all other positions for all $j = 1, \ldots, n_i$. $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_M)$ denotes the parameter vector with $\mu_i = \mu + \alpha_i$. The comparisons of interest (3.2) can be specified as linear functions of parameters from the linear model

$$\mathrm{H}^0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$$

using the Tukey-type contrast matrix

$$\boldsymbol{C} = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ & & & & \vdots & & & \\ -1 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ & & & & \vdots & & & \\ & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{k \times M}, \tag{3.4}$$

and $k = (M(M+1))/2$ being the number of all pairwise comparisons. Each row of the matrix $\boldsymbol{C}$ corresponds to one of the partial hypotheses $\mathrm{H}_{i\imath}^0$. The right hand side of the hypotheses is specified as $\boldsymbol{a} = (0, \ldots, 0) \in \mathbb{R}^k$. Further pairwise comparisons procedures like Dunnett's many-to-one comparisons can be specified by a Dunnett-type contrast matrix $\boldsymbol{C}$, as described in Section 2.5.

### 3.1.2 Estimation

Estimates of the parameters and the associated covariance matrix are obtained from the linear model representation. Let $\boldsymbol{X}$ be the design matrix of model (3.3) whose rows are the vectors $\boldsymbol{x}_{ij}$, $i = 1, \ldots, M$, $j = 1, \ldots, n_i$, $\boldsymbol{y} = (y_{11}, \ldots y_{Mn_M})$ the vector of the dependent measurements, and $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots \varepsilon_{Mn_M})$ the vector of random errors with $\mathbb{E}(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}) = \boldsymbol{\Phi} = \text{diag}(\sigma_i^2)$. The OLS estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \in \mathbb{R}^M$$

is best linear unbiased and asymptotically normal

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Phi} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \in \mathbb{R}^{M \times M} \tag{3.5}$$

under both homoscedasticity and heteroscedasticity (Eicker, 1963). The OLS covariance estimation

$$\text{OLSCM} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{n_i} e_{ij}^2}{\sum_{i=1}^{M} n_i - M} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

with least squares residuals $e_{ij} = \boldsymbol{x}_{ij}^\top \hat{\boldsymbol{\beta}}$ is consistent when the assumption of homoscedastic errors holds, but inconsistent under heteroscedastic errors (White, 1980). Several covariance estimation matrices that are consistent under heteroscedasticity of unknown form have been proposed. In the following, we consider the covariance estimates HC3 (Davidson and MacKinnon, 1993), HC4 (Cribato-Neto, 2004), and HC4m (Cribato-Neto and da Silva, 2011). All of them are modifications of the first suggestion for a heteroscedastic-constistent covariance estimation by White (1980).
The covariance estimation matrices plug an estimate

$$\hat{\boldsymbol{\Phi}} = \boldsymbol{E} \cdot \hat{\boldsymbol{\Omega}}$$

for $\boldsymbol{\Phi}$ in the equation of the covariance (3.5) with $\hat{\boldsymbol{\Omega}} = \text{diag}\{e_{11}, \ldots, e_{Mn_M}\}$ and $\boldsymbol{E}$ differing between the different heteroscedastic-consistent covariance estimations. The calculation of $\boldsymbol{E}$ is based on the diagonal elements of the hat matrix

$$\boldsymbol{H} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$$

which measures the leverage of the corresponding observations in the linear model. For the special case of an ANOVA model the diagonal elements of $\boldsymbol{H}$ only depend on the group size of the corresponding observation and equal $1/n_i$ (Hoaglin and Welsch, 1978).

In simulations by Long and Ervin (2000) the estimation HC3 with

$$\boldsymbol{E}_{\text{HC3}} = \text{diag}\left\{ \frac{1}{(1 - 1/n_i)^2} \right\}$$

performed well under finite sample sizes. The HC4 estimator uses

$$\boldsymbol{E}_{\mathrm{HC4}} = \mathrm{diag}\left\{\frac{1}{(1-1/n_i)^{\delta_i}}\right\} \quad \text{with} \quad \delta_i = \min\left\{4, \frac{n/n_i}{M}\right\},$$

which leads to a stronger inflation of the residuals belonging to smaller groups. A modified version HC4m uses

$$\boldsymbol{E}_{\mathrm{HC4m}} = \mathrm{diag}\left\{\frac{1}{(1-1/n_i)^{\delta_i}}\right\} \quad \text{with} \quad \delta_i = \min\left\{1, \frac{n/n_i}{M}\right\} + \min\left\{1.5, \frac{n/n_i}{M}\right\}.$$

## 3.2 Behavior of Tukey-Type Comparisons Under Heteroscedasticity

### 3.2.1 Simulation Setup

For all pairwise comparisons of group means the familywise error rate and the power properties of simultaneous inference using the covariance estimations HC3, HC4, and HC4m were estimated and compared to the Tukey-Kramer method. The Tukey-Kramer method compares samples of unequal sizes all-pairwise based on the studentized range distribution under the assumption of normal data and equal variances among all groups (Kramer, 1956).

We considered unbalanced one-way ANOVA models with $M = 4$ groups under homoscedasticity, under heteroscedasticity with smaller variances in the smaller groups and vice versa both for normal and non-normal, right-skewed data. For the classical procedures, these special conditions of positive or negative pairing of group sizes and variances typically lead to conservative or liberal results, respectively.

A: $n_1 < n_2 < n_3 < n_4$ and $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$, normal data.

B: $n_1 < n_2 < n_3 < n_4$ and $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$, skewed data.

C: $n_1 < n_2 < n_3 < n_4$ and $\sigma_1 < \sigma_2 < \sigma_3 < \sigma_4$, normal data.

D: $n_1 < n_2 < n_3 < n_4$ and $\sigma_1 < \sigma_2 < \sigma_3 < \sigma_4$, skewed data.

E: $n_1 < n_2 < n_3 < n_4$ and $\sigma_1 > \sigma_2 > \sigma_3 > \sigma_4$, normal data.

F: $n_1 < n_2 < n_3 < n_4$ and $\sigma_1 > \sigma_2 > \sigma_3 > \sigma_4$, skewed data.

Total sample sizes of $N = 60, 120, 180, 240$ were considered with the $N$ observations unbalancedly distributed to the four groups. The number of observations $n_i$ for each group $i = 1, \ldots 4$ were defined as $n_i = n + 0.2 \cdot i \cdot n$, $i = 1, \ldots, 4$, $n = 10, 20, 30, 40$, leading to $\sum_i n_i = N$. For estimation of the FWER all group means were set equal $\mu_i = 2$, $i = 1, \ldots, 4$. In models A, C, and E the random errors were independently normally distributed $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ with group specific standard deviations $\sigma_i$. In models B, D, and F the random errors were independently simulated from a shifted and

scaled Beta distribution with mean 0 and group specific standard deviations $\sigma_i$. Standard deviations $\sigma = (\sigma_1, \ldots, \sigma_4)$ were chosen as $\sigma = (2, 2, 2, 2)$ in model A and B, $\sigma = (1.25, 1.75, 2.25, 2.75)$ in models C and D, and $\sigma = (2.75, 2.25, 1.75, 1.25)$ in models E and F.

41,000 datasets of size $N = \sum_i n_i$ were simulated according to the considered models A to F. In each dataset all pairwise comparisons of the group means were tested by the simultaneous inference approach using the covariance estimations HC3, HC4, and HC4m, and by the Tukey-Kramer method. 41 estimates of the familywise error rate were calculated from the proportion among 1,000 datasets, in which at least one true partial hypothesis was falsely rejected.
To investigate the power of the procedures the means of groups 2 to 4 ($\mu_2, \mu_3$ and $\mu_4$) were kept equal while the mean of the first group $\mu_1$ was chosen differently. Thus, the pairwise comparisons of $\mu_1$ with each of the three other means were false. For each of these false partial hypotheses the individual power of simultaneous tests was estimated by the proportion of correctly rejected partial hypotheses among 1,000 datasets for increasing distances between $\mu_1$ and $\mu_i$, $i = 2, 3, 4$. 41 values of distances $\mu_1 - \mu_i$, $i = 2, 3, 4$, were considered. The nominal level $\alpha$ was 0.05.

### 3.2.2 Simulation Results

#### Familywise Error Rate

The distribution of the estimated familywise error rates for all pairwise comparisons of group means for the Tukey-Kramer method and the simultaneous inference approaches using the heteroscedastic-consistent covariance estimations HC3, HC4, and HC4m are illustrated in Figures 3.1, 3.2, and 3.3. The boxplot for each setting is calculated from the 41 estimated values.
In all settings A to F the simultaneous tests using HC4 yield an estimated familywise error rate more distant from the nominal level $\alpha = 0.05$ compared to the tests using HC3 and HC4m. In the models with equal variances in all groups (models A and B) the estimated familywise error rate is close to 0.05 for the Tukey-Kramer method and for the testing procedures using HC3 and HC4m. With unequal variances and higher variances in the larger groups (models B and C), the Tukey-Kramer method is conservative while the estimated familywise error rate of the tests using HC3 or HC4m is close to 0.05 already for a total sample size of $N = 60$ for normal data, and $N = 120$ for skewed data. In the situation with higher variances in the smaller groups for both normal or skewed data (models E and F), the Tukey-Kramer method is very liberal. For a total sample size of $N = 60$ the simultaneous tests using the consistent covariance estimations HC3 and HC4m are liberal as well, but less than the Tukey-Kramer method. With increasing total sample sizes their estimated familywise error rate approaches the nominal level well.
The similar results regarding FWER and power for the covariances HC3 and HC4m result from similar values in the weight matrices $\boldsymbol{E}_{\mathrm{HC3}}$ and $\boldsymbol{E}_{\mathrm{HC4m}}$ for the inspected number of groups and degrees of imbalance.

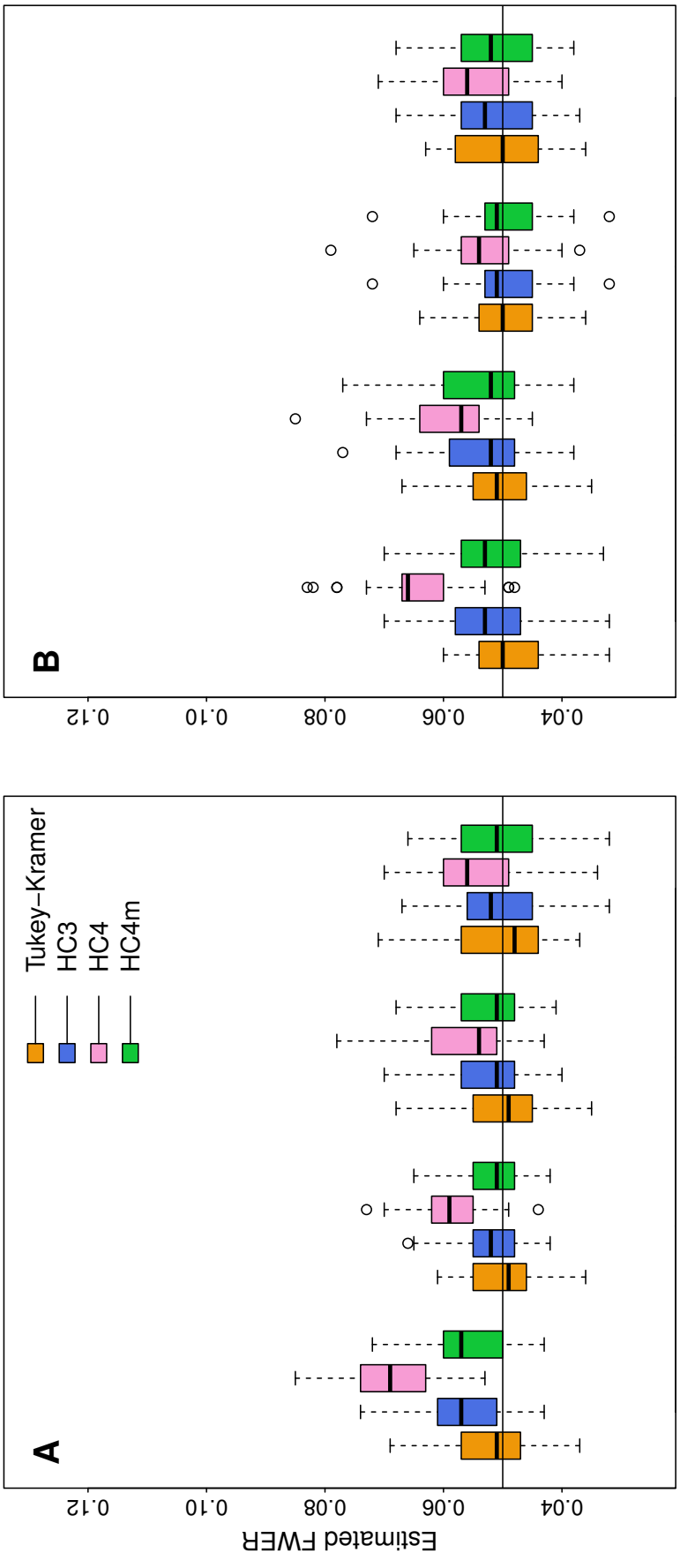


Figure 3.1: Estimated familywise error rate of the Tukey-Kramer method and of the simultaneous tests using the heteroscedastic consistent covariance estimations HC3, HC4, and HC4m for all pairwise comparisons of means of normal samples (A) and skewed samples (B) under homoscedasticity. The total number of observations $N$ was unbalancedly distributed among the four groups.

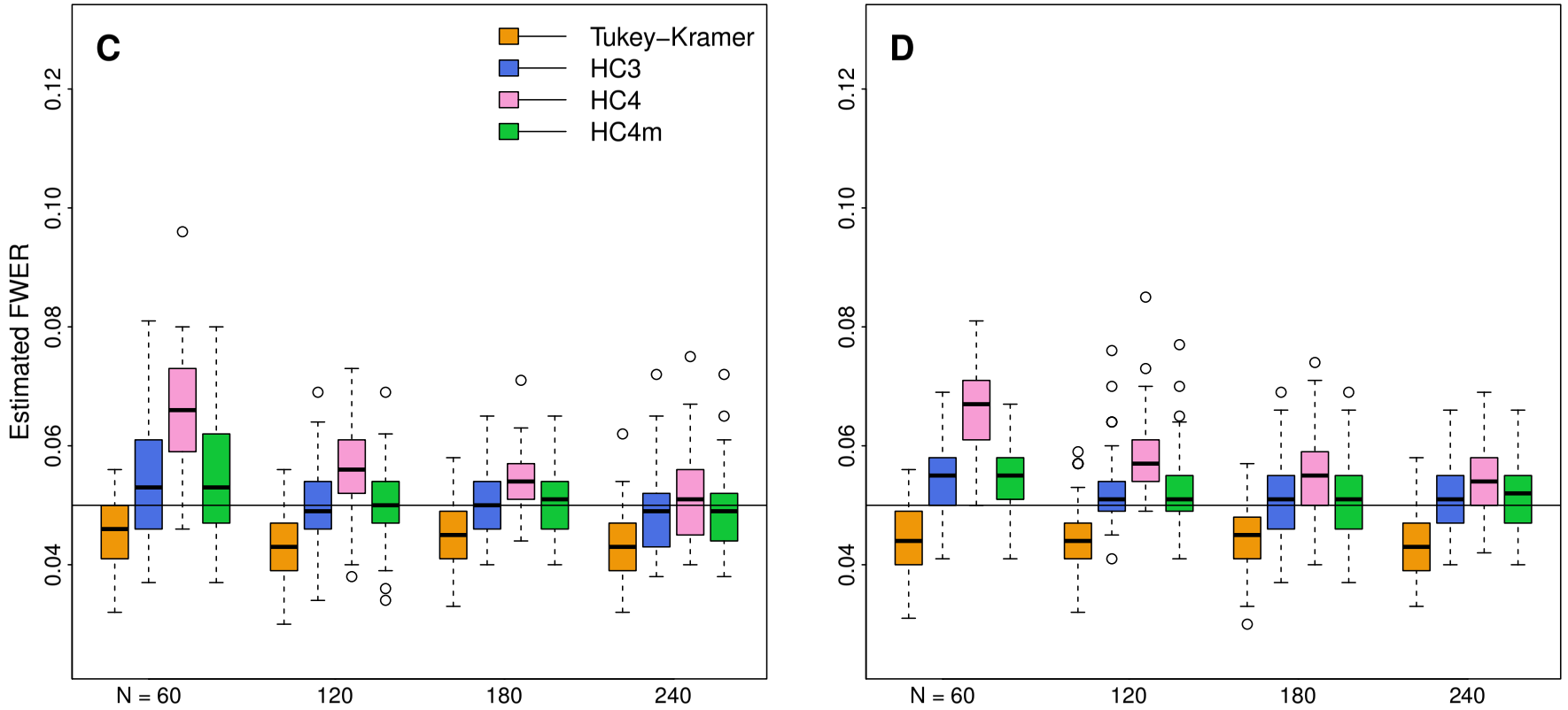


Figure 3.2: Estimated familywise error rate of the Tukey-Kramer method and of the simultaneous tests using the heteroscedastic consistent covariance estimations HC3, HC4, HC4m for all pairwise comparisons of means of normal samples (C) and skewed samples (D) under heteroscedasticity with smaller variances in the smaller groups. The total number of observations $N$ was unbalancedly distributed among the four groups.

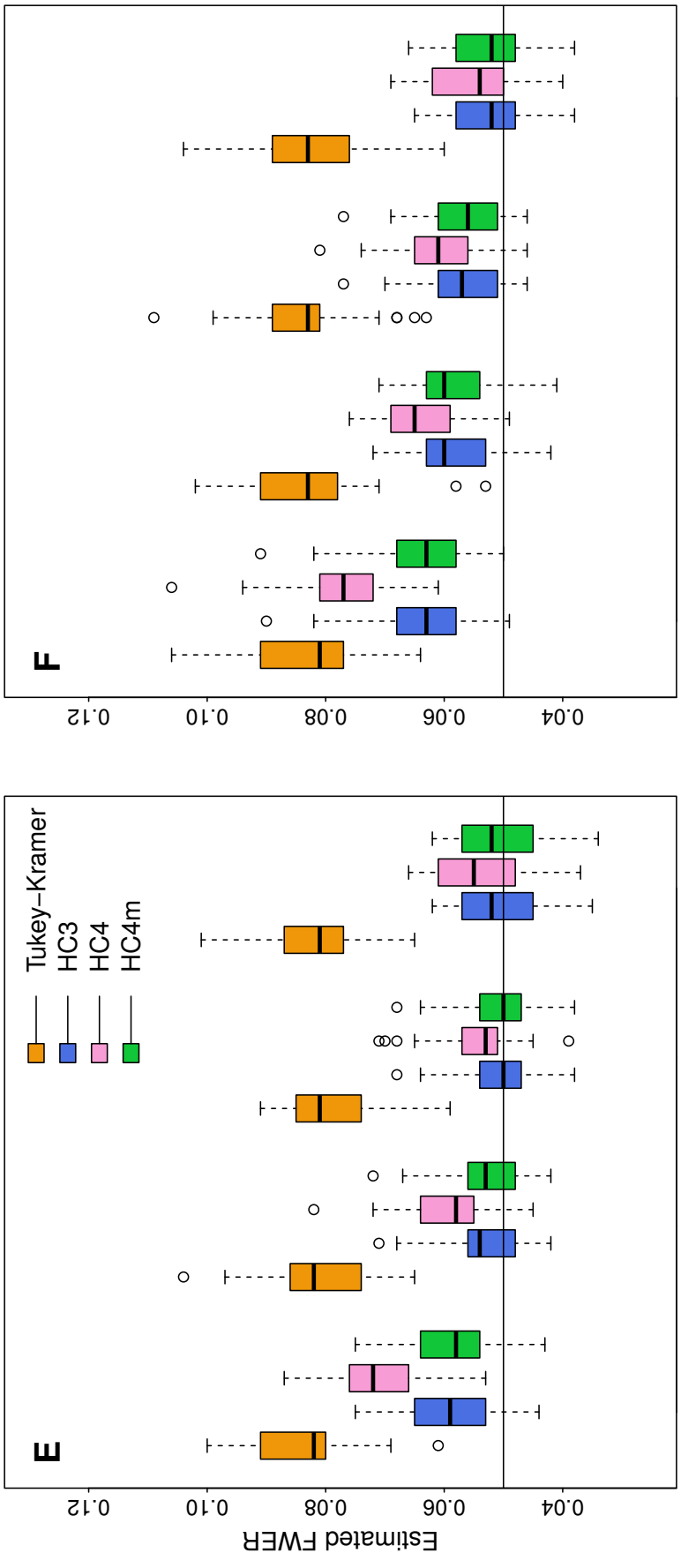


Figure 3.3: Estimated familywise error rate of the Tukey-Kramer method and of the simultaneous tests using the heteroscedastic consistent covariance estimations HC3, HC4, HC4m for all pairwise comparisons of means of normal samples (E) and skewed samples (F) under heteroscedasticity with smaller variances in the larger groups. The total number of observations $N$ was unbalancedly distributed among the four groups.

### Power

Since the heteroscedastic covariance estimations HC3 and HC4m performed better than HC4 in terms of the familywise error rate, the approach using HC4 was dropped from the power analyses. Figures 3.4, 3.5, and 3.6 show the individual power curves of the Tukey-Kramer method and of the simultaneous tests using HC3 for models A to F for the three pairwise comparisons of group means $\mu_i$, $i = 2, 3, 4$, with $\mu_1$, when the mean of the first group differs from the others. The power results of the tests using HC4m are comparable to the results for HC3 and are not displayed.

The power properties of neither testing procedure differs between settings with normal data compared to skewed data under homoscedasticity and both scenarios of heteroscedasticity. Under homoscedasticity (models A and B) the individual power curves of both procedures are almost identical for equivalent sample size $N$ (see Figure 3.4). In the heteroscedastic settings with positive pairing of variances and group sizes the power of the simultaneous tests using HC3 achieve a higher power compared to the Tukey-Kramer method (see Figure 3.5). In accordance with the conservative character of the Tukey-Kramer test its power is lower in the heteroscedastic settings C and D than in the homoscedastic settings. While the FWER of the simultaneous tests using HC3 is comparable between the heteroscedastic settings C and D and the homoscedastic settings the power is slightly higher under this kind of heteroscedasticity. In the heteroscedastic settings with positive pairing of variances and group sizes the individual power for all false hypotheses is higher for the Tukey-Kramer method compared to the approach using HC3 (see Figure 3.6), but yet the former cannot be recommended because the familywise error rate is not controlled. The power of the simultaneous tests using HC3 is lower in the settings with smaller variances in the larger groups compared to the inverse heteroscedastic settings.

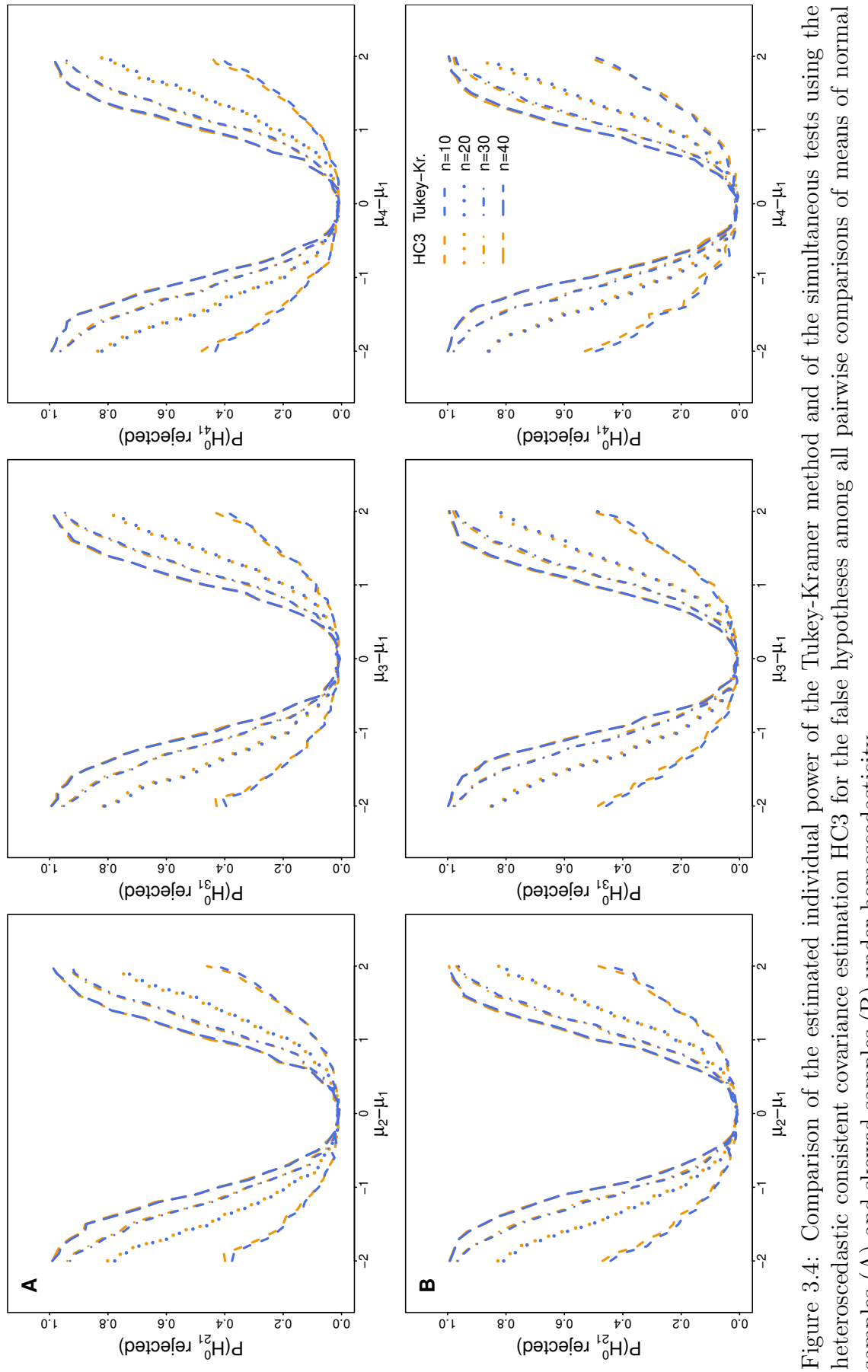


Figure 3.4: Comparison of the estimated individual power of the Tukey-Kramer method and of the simultaneous tests using the heteroscedastic consistent covariance estimation HC3 for the false hypotheses among all pairwise comparisons of means of normal samples (A) and skewed samples (B) under homoscedasticity.

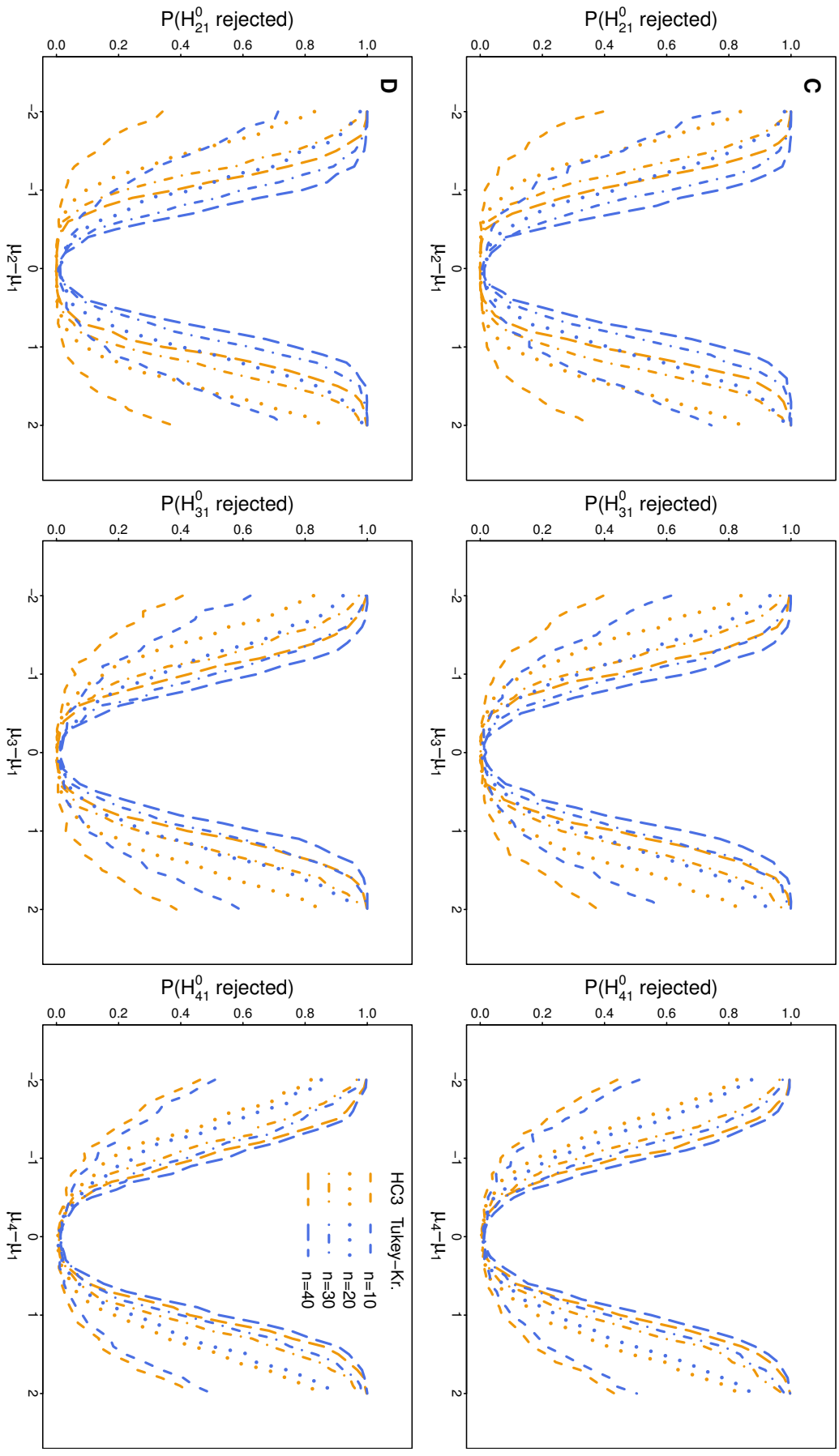


Figure 3.5: Comparison of the estimated individual power of the Tukey-Kramer method and of the simultaneous tests using the heteroscedastic consistent covariance estimation HC3 for the false hypotheses among all pairwise comparisons of means of normal samples (C) and skewed samples (D) under heteroscedasticity with smaller variances in the smaller groups.

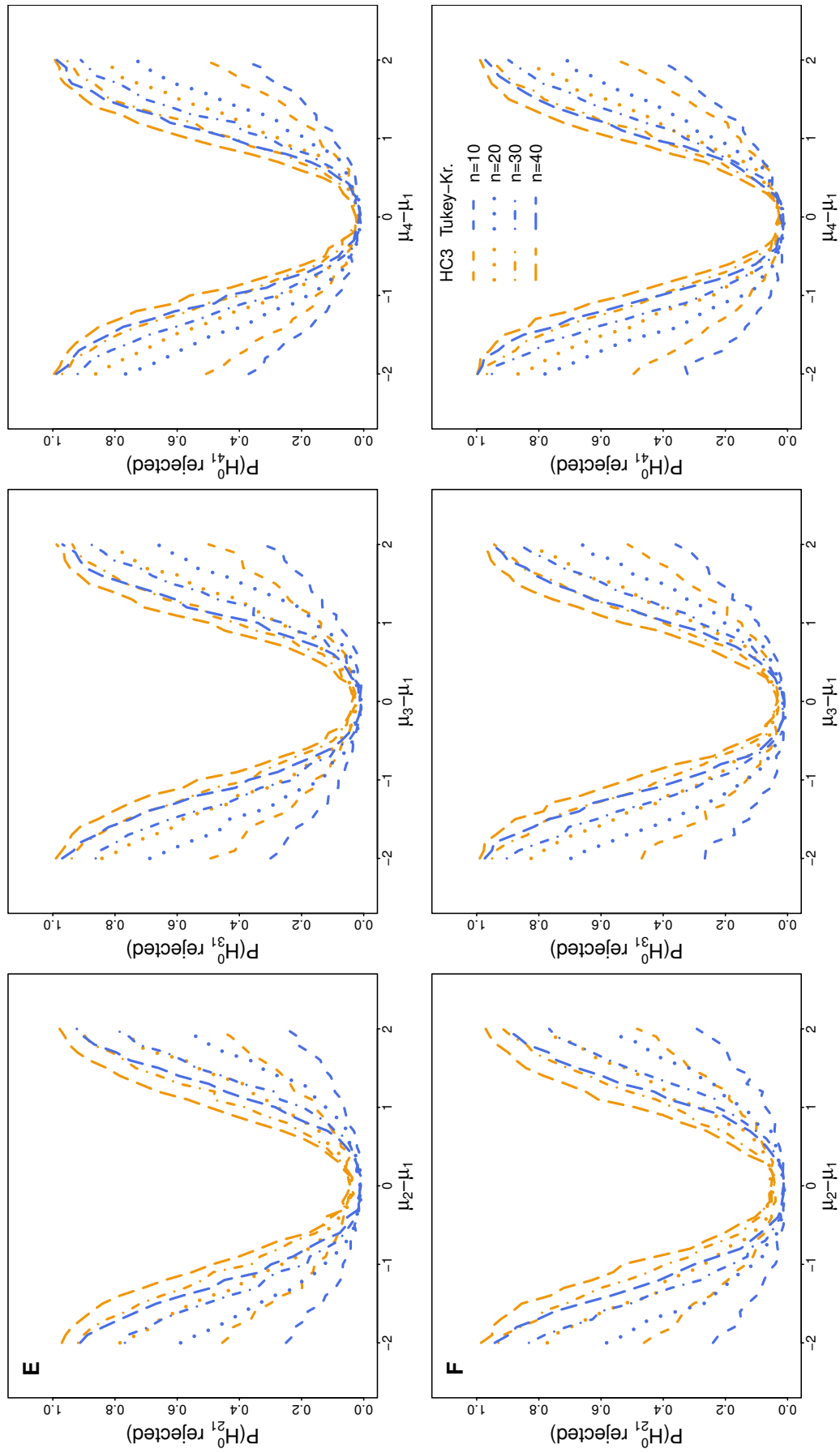


Figure 3.6: Comparison of the estimated individual power of the Tukey-Kramer method and of the simultaneous tests using the heteroscedastic consistent covariance estimation HC3 for the false hypotheses among all pairwise comparisons of means of normal samples (E) and skewed samples (F) under heteroscedasticity with smaller variances in the larger groups.

## 3.3 Comparisons of Fatty Acid Phenotypes of *Bacillus Simplex* Putative Ecotypes

The *Bacillus simplex* population from "Evolution Canyons" I and II in Israel has recently developed to a model study of bacterial adaptation and speciation under heterogeneous environmental conditions (Sikorski and Nevo, 2005). These two canyons represent similar ecological sites 40 km apart in which the orientation of the sun yields a strong sun-exposed and hot 'African' south-facing slope versus a rather cooler and mesic-lush 'European' north-facing slope within a distance of only 50–400 m. Phylogenetically, based on DNA sequences, the *B. simplex* population splits into two major groups GL1 and GL2. Interestingly, within each GL1 and GL2, further phylogenetic groups (or so called 'putative ecotypes') were observed which show a clear preference for either slope type (Sikorski and Nevo, 2005; Sikorski et al., 2008b). Putative ecotypes (PE) we regard as phylogenetic lineages whose members are adapted to specific ecological conditions (Cohan and Perry, 2007; Koeppel et al., 2008). Whereas GL2 is composed of only PE1 and PE2, GL1 is made up of multiple PE (PE3–PE9) (Koeppel et al., 2008; Sikorski et al., 2008b). The bacteria's characteristic slope type preference might be explained by physiological properties (phenotypes) such as temperature stress related phenotypes as a putative evolutionary adaptive response to the different temperatures on both slopes. For example, the physical integrity of the cell membrane at different temperatures is crucial for cell survival and in turn it crucially depends on the fatty acid (FA) composition of the cell membrane. This was the motivation for a recent study on the contents of high- and low-temperature-tolerance-providing fatty acids (FAs) of the *B. simplex* ecotypes (Sikorski et al., 2008a). However, as the methods for the genetic characterization were improved in the meantime, leading to a re-shuffling of individuals into different groups (see also Table 3 of the supplemental material of Koeppel et al., 2008) and as the former fatty acid data were analyzed using the classical non-robust statistical tools (Sikorski et al., 2008a) it seems worthwhile to reanalyze the experiment using the simultaneous inference approach employing heteroscedastic-consistent covariance estimation. We focus specifically on the multiple ecotypes PE3 to PE9 from GL1 (we exclude PE8, as this ecotype is represented by only two bacterial strains).

Heteroscedasticity among the PE is assessed visually by boxplots, which illustrate the distribution of the FAs for the six PE. Analyses are conducted both with the Tukey-Kramer method assuming homogeneous variances and with simultaneous approaches accounting for heteroscedasticity to investigate in which way wrong conclusions can be drawn when heterogeneous variances are ignored. We compute simultaneous confidence intervals for all pairwise differences of group means to investigate which pairs of PE differ significantly concerning a specific growth condition of the bacteria (Sikorski et al., 2008a).

Figure 3.7 shows the distributions of high- and low-temperature-tolerance-providing FAs in six PE of *B. simplex* (PE3 – PE9) for six different experimental conditions (Figures 3.7a–f).

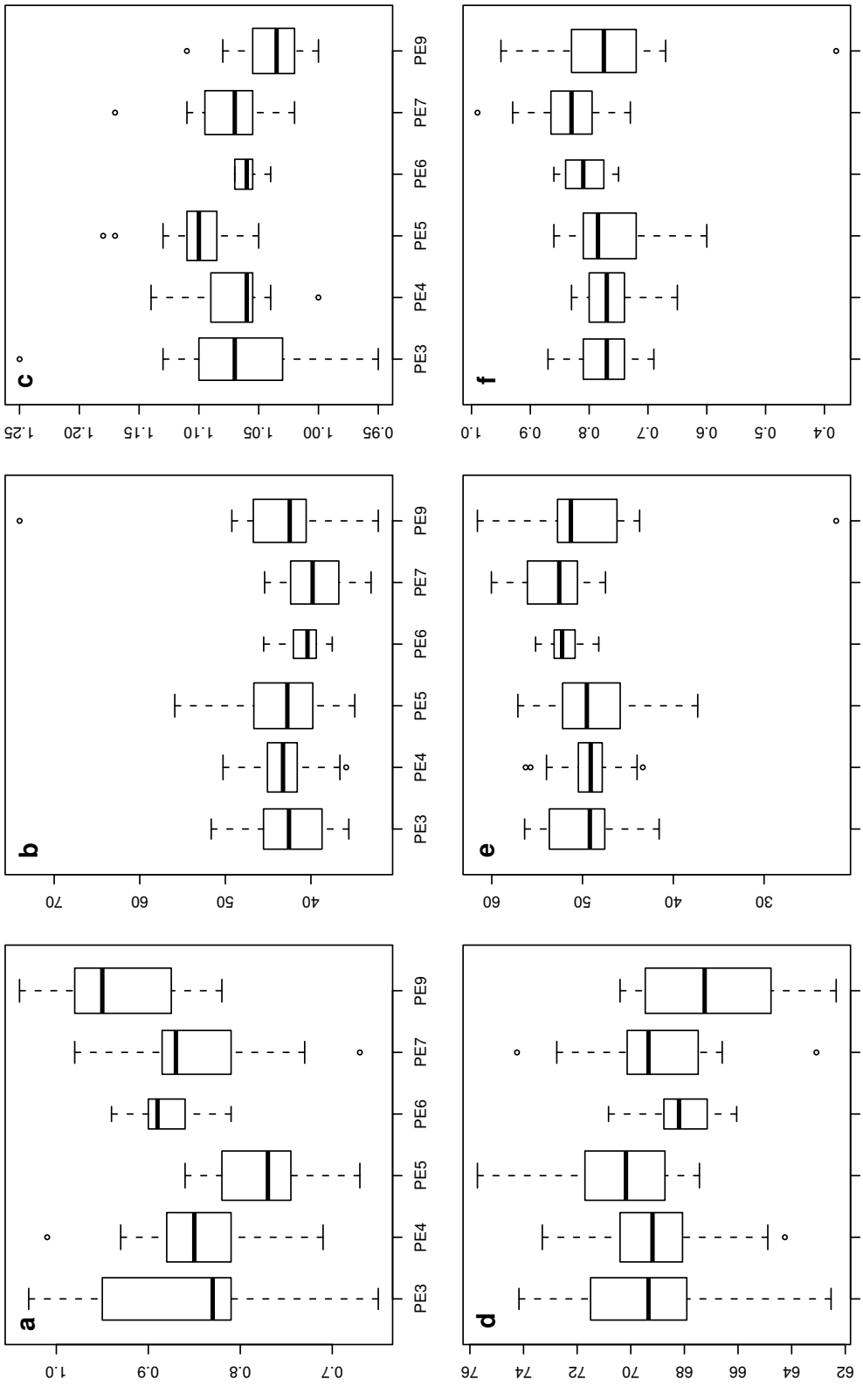


Figure 3.7: Distribution of the fatty acid content in six lineages (putative ecotypes, PE) of *B. simplex* for six different experimental conditions (a to f). Strains were grown on Trypticase Soy Broth Agar (Difco) for 24 hours at different temperatures. Fig. a and b sum up the high-temperature tolerance providing iso-branched FAs. Fig. c to f sum up the cold-temperature tolerance providing anteiso-branched and unsaturated FA. Further experimental details are described in Sikorski et al. (2008a).

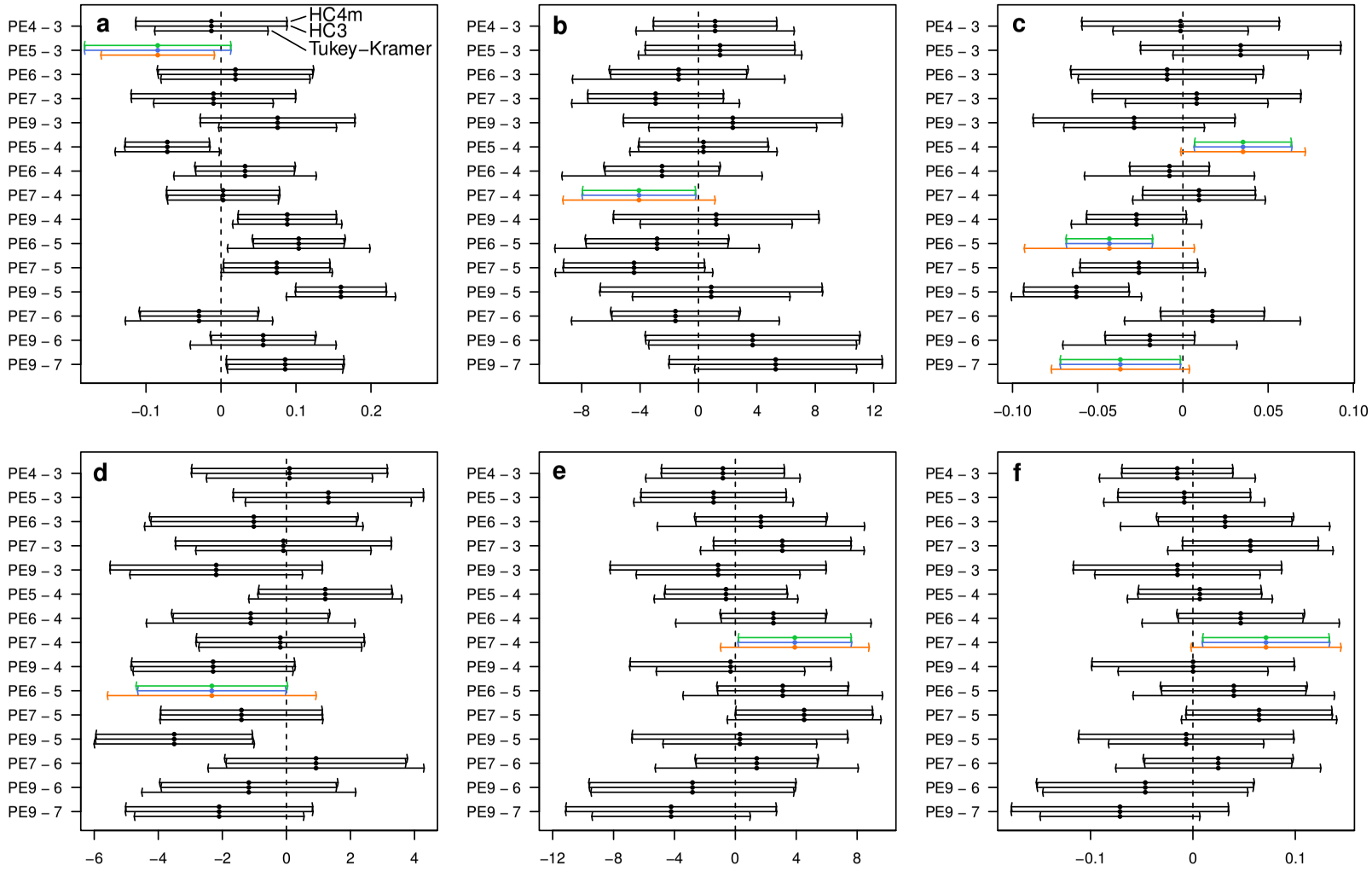


Figure 3.8: Simultaneous confidence intervals for all pairwise comparisons of group means for six different experimental conditions a-f. Intervals are computed by the approaches accounting for heteroscedasticity using the heteroscedastic consistent covariance estimations HC3 and HC4m, respectively, and by the Tukey-Kramer method assuming homoscedasticity. The colored confidence intervals indicate the pairwise comparisons for which the decision of significant difference of the associated group means differs between the test procedures.

Variances differ considerably between the lineages within each experimental condition. Thus, the validity of the results of the Tukey-Kramer method might be in question and attention should be drawn to the results of simultaneous inference employing the covariance estimations HC3 and HC4m, which performed well in the simulations. Results of the Tukey-Kramer method are presented in addition to the heteroscedastic-consistent methods to show the extent of differences in the results.

The widths of the simultaneous confidence intervals based on the heteroscedastic-consistent covariance estimations HC3 and HC4m do not differ between them in any comparison of strains, but are noticeably different from the intervals calculated by the Tukey-Kramer method, either narrower or wider (see Figure 3.8).

Two PE are considered significantly different concerning their fatty acid content if the associated simultaneous confidence interval does not include zero. For several comparisons the decision of significant difference depends on the method chosen (colored simultaneous confidence intervals). When heterogeneous variances are neglected, a significant difference in the lineages PE3 and PE5 is found concerning the FAs (Figure 3.8a), which is not present when heteroscedasticity is accounted for. For the other FAs (Figures 3.8b–f) significantly differing lineages of *B. simplex* are not detected, when heteroscedasticity is ignored.

## 3.4 Summary

We investigated the behavior of the simultaneous inference procedure by Hothorn et al. (2008) for all pairwise comparisons of means in samples with heterogeneous variances. Heteroscedastic consistent sandwich matrices were used as estimators for the covariance matrix. The approach is applicable for multiple comparisons under potential heteroscedasticity in balanced or unbalanced designs with arbitrary error distribution. Usage of the covariance estimations HC3 and HC4m revealed better performance than HC4 in terms of the familywise error rate. For the simultaneous tests using HC3 or HC4m the familywise error rate is controlled already for relatively small samples in unbalanced designs with normal or skewed samples and different kinds of pairing of group sizes and variance. By contrast, the Tukey-Kramer method, which assumes homoscedasticity, is either conservative or liberal depending of the relation between sample sizes and heteroscedasticity and increasing the total sample size does not eliminate the departure of the FWER from the nominal level. In settings where the Tukey-Kramer method does not show increased false positive rates, the approaches using HC3 or HC4m achieve higher power to detect existing group differences.
In conclusion, the simultaneous inference procedure by Hothorn et al. (2008) using a heteroscedastic consistent covariance estimation performs well for multiple comparisons of means in presence of unequal variances.

# Chapter 4

# Multiple Comparisons in Survival Models

The design of medical or biological studies often includes several treatment groups and a control group. In this chapter the behavior of simultaneous inference in survival models using the procedure by Hothorn et al. (2008) is studied for multiple comparisons of survival in several treated groups with a control. The evaluation of simultaneous confidence intervals for Dunnett-type contrasts in the frailty Cox model with the objective of comparing several experimental therapies with a standard therapy in multicenter clinical trials is based on Herberich and Hothorn (2012b). The evaluation of simultaneous confidence intervals for Williams-type contrasts in the Cox proportional hazards model and in the accelerated failure time model with the objective of trend analyses in toxicological studies extends the results in Herberich and Hothorn (2012a).

Clinical trials are often conducted with the objective of comparing the survival of patients undergoing one of several experimental therapies with that of one standard therapy. Examples are a multicenter trial of the Eastern Cooperative Oncology Group, in which the combined chemotherapies gemcitabine plus carboplatin and gemcitabine plus paclitaxel were compared with the standard regimen of paclitaxel plus carboplatin in patients with advanced or metastatic non-small-cell lung cancer (Treat et al., 2010) and a multicenter trial of the Alpha Oncology Research Network, in which a fixed-dose rate of gemcitabine and a combination of gemcitabine plus oxaliplatin were compared with the standard treatment of single-agent gemcitabine in patients suffering from pancreatic cancer (Poplin et al., 2009). A similar question was addressed in a multicenter trial of the the German CML Study Group, in which the survival of patients with chronic myelogenous leukemia (CML) after treatment with the cytokine Interferon-$\alpha$ was compared with that after the two conventional chemotherapies busulfan and hydroxyurea (Hehlmann et al., 1994).

These Dunnett-type comparisons lead to an inflated probability of false-positive results

if no adjustment for multiplicity is applied. For Gaussian or binomial data statistical tools for many-to-one comparisons are provided for a variety of settings (Dunnett, 1955; Dunnett and Tamhane, 1992; Chuang-Stein and Tong, 1995; Schaarschmidt, Biesheuvel, and Hothorn, 2009; Klingenberg, 2010). In contrast, the comparison of several groups with regard to time of survival has received less attention in the statistical literature. The generalization of the log-rank test to multiple groups only allows for a global comparison of several survival curves (see e.g. Kulathinal and Gasbarra, 2002). Using several two-sample log-rank tests to investigate which particular groups differ with regard to time to event requires multiplicity adjustment. The standard procedure also used in the clinical trials mentioned above, is the Bonferroni correction. It does not take dependencies among the test statistics into account leading to conservative results and hence less power to detect clinically relevant differences. Further adjustments that account for correlation among the tests have been proposed (Chakraborti and Desu, 1991; Chen, 2000; Logan, Wang, and Zhang, 2005). However, log-rank-based tests neglect additional covariates and can only account via stratification for patients being recruited in different study centers. Besides, log-rank tests only address statistical significance and do not provide an interpretable measure for decisions on the clinical relevance. Simultaneous confidence intervals for Dunnett-type contrasts of treatment effects in survival models can be obtained by the simultaneous inference procedure for general parametric models by Hothorn et al. (2008) and used for multiple comparisons of several experimental treatments with a control treatment. Adjustment for further covariates such as sex and age is possible by extending the predictor of the survival model. In case of clustered data, e.g. due to recruitment of patients in different study centers, frailty models can be used. They account for potential correlation of subjects belonging to the same cluster using random effects.

A further application of simultaneous inference in survival models are toxicity studies, which compare mortality, among other endpoints, between dosages of a drug or compound to a negative control. An example is the study NTP-120 of the National Toxicology Program (NTP) where the effects of increasing dosages of piperonyl butoxide, an organic compound used as pesticide synergist, are investigated in mice and rats (National Toxicology Program (U.S.), 1979). Common designs use balanced samples with one negative control receiving no substance and a relatively small number of treatment groups receiving increasing dosages of the drug or compound. The current approach for the analysis of survival/mortality is Tarone's lifetable test (Tarone, 1975), which tests for a linear trend (National Toxicology Program (U.S.), 2012). However, this approach lacks power to detect non-linear trend shapes. Williams (1971, 1972) introduced an approach for comparisons of dose groups to control under total order restriction, i.e. under the assumption of a strictly monotone dose-response relationship, for normal data. Bretz and Hothorn (2000) generalized this approach to a multiple contrast test where each of the so-called Williams-type contrasts defines a dose-response shape. Trend analyses for survival endpoints can be performed using the approach by Hothorn et al. (2008) for multiple comparisons of Williams-type contrasts in survival models. Simultaneous confidence intervals for these contrasts can be transformed to simultaneous confidence intervals for hazard ratios in the Cox proportional hazards (PH) model and to simultaneous confidence intervals for survival time ratios in the accelerated failure time (AFT) model. Simultane-

ous confidence intervals for these effect sizes allow for interpretation of a trend in terms of both statistical significance and biological relevance.

This chapter first establishes the validity of the procedure by Hothorn et al. (2008) for survival models. In the simulations special attention is being paid to small sample sizes as these are often required for ethical reasons in clinical trials and toxicity studies. The joint coverage probability of simultaneous confidence intervals for Dunnett-type hazard ratios is estimated in the frailty Cox model with the objective of comparing several experimental treatments with a standard treatment in multicenter clinical trials. In the second part of this chapter, the joint coverage probability and power properties of simultaneous confidence intervals for Williams-type comparisons are studied in the Cox proportional hazards (PH) model and in the accelerated failure time (AFT) model. These simultaneous confidence intervals allow to analyze the effects of increasing dosages on mortality in long-term carcinogenicity studies.
The applicability of the method for many-to-one comparisons is illustrated by a reanalysis of the data from the German CML Study Group reported by Hehlmann et al. (1994), which compares the standard treatment for CML to two different experimental treatments, with adjustment for covariates and taking the multicenter structure into account. The trend approach is applied to a reanalysis of mortality in the NTP-120 study of piperonyl butoxide in female mice.

## 4.1 Simultaneous Inference in Survival Models

### 4.1.1 Models and Hypotheses

The semiparametric Cox proportional hazards model describes the hazard for subject $j$ as

$$\lambda(t|\boldsymbol{x}_j) = \lambda_0(t)\, \exp\left(\boldsymbol{x}_j^\top \boldsymbol{\beta}\right), \quad j = 1, \ldots, n.$$

$\lambda_0(t)$ denotes the baseline hazard rate at time $t$ and is assumed to be identical for all individuals, the vector $\boldsymbol{x}_j$ includes the covariates of the $j$th individual, and $\boldsymbol{\beta}$ is the associated vector of regression coefficients.

The frailty Cox model specifies the hazard for subject $j$ belonging to the $i$th of $I$ clusters as

$$\begin{aligned} \lambda(t|\boldsymbol{x}_{ij}) &= \lambda_0(t)\, \exp\left(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \gamma_i\right) \\ &= \lambda_0(t)\, \nu_i\, \exp\left(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}\right), \quad i = 1, \ldots, I,\ j = 1, \ldots, n_i. \end{aligned}$$

$\lambda_0(t)$ denotes the baseline hazard rate at time $t$ and is assumed to be identical for all individuals, the vector $\boldsymbol{x}_{ij}$ includes the covariates of the $j$th individual in the $i$th cluster, and $\boldsymbol{\beta}$ is the associated vector of regression coefficients. The frailties $\nu_i = \exp(\gamma_i)$ describe the excess risk for cluster $i$ and are assumed to be independent and identically distributed from a gamma or log-normal distribution with variance $\theta$ and mean 1.

The accelerated failure time (AFT) model provides a parametric approach to modeling time of survival in several treatment groups in the case of non-proportional hazards. In the log-linear AFT model

$$\log(t_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \sigma\epsilon_i \quad \Leftrightarrow \quad t_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \cdot \exp(\sigma\epsilon_i), \quad i = 1, \ldots, n,$$

the exponentiated effects act multiplicatively in terms of the time $t$, and additively on terms of $\log t$. The random errors $\epsilon_i$ are assumed to have a particular distribution with scale parameter $\sigma$ corresponding to a particular assumed distribution for the survival times (Kalbfleisch and Prentice, 2002).

Let the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M, \beta_{M+1}, \ldots, \beta_p) \in \mathbb{R}^p$ contain the effects $\beta_m$, $m = 1, \ldots, M$, of a grouping variable with constraint $\sum_{m=1}^{M} \beta_m = 0$ and the effects $\beta_{M+1}, \ldots \beta_p$ of $p - M$ further covariates, if applicable. It is of interest to compare the effects on survival between the $M$ levels of the grouping variable. For two individuals who differ only in the group membership (e.g., experimental vs. standard treatment), their predicted hazard rate from the Cox PH model or frailty Cox model will differ by $\exp(\beta_2 - \beta_1)$, which can be considered a hazard ratio between the predicted hazard for a member of group 2 and that for a member of group 1, holding all other variables constant. In the AFT model, $\exp(\beta_2 - \beta_1)$ corresponds to the survival time ratios between the predicted survival for a member of group 2 and that for a member of group 1, holding all other variables constant. Confidence intervals for hazard ratios or survival time ratios, respectively, can provide information on clinical or biological relevance of differences between groups in addition to conclusions about statistical significance.
Dunnett-type differences of group effects for comparisons of several experimental groups with a control group can be set up as linear functions of the parameter vector $\boldsymbol{\beta}$ by the contrast matrix

$$\boldsymbol{C} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & \cdots \\ -1 & 0 & 1 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & \cdots & 1 & 0 & \cdots \end{pmatrix} \in \mathbb{R}^{M-1 \times p},$$

where each row of the linear combination $\boldsymbol{C\beta}$ corresponds to one of the pairwise comparisons $\beta_m - \beta_1$, $m = 2, \ldots, M$, i.e., to a log hazard ratio for the Cox PH model and to a log survival time ratio for the AFT model, respectively. Group 1 denotes the control group and groups $m = 2, \ldots, M$ denote the experimental groups.

For analysis of a trend in survival/mortality the null hypothesis of no difference between treatment effects against a monotone ordered alternative can be assessed by Williams-type contrasts of group effects, which are specified by the contrast matrix

$$\boldsymbol{C} = \begin{pmatrix} -1 & 0 & \cdots & 0 & 0 & 1 & 0\cdots \\ -1 & 0 & \cdots & 0 & \frac{n_{M-1}}{n_{M-1}+n_M} & \frac{n_M}{n_{M-1}+n_M} & 0\cdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ -1 & \frac{n_2}{\sum_{m=2}^M n_m} & \cdots & \frac{n_{M-2}}{\sum_{m=2}^M n_m} & \frac{n_{M-1}}{\sum_{m=2}^M n_m} & \frac{n_M}{\sum_{m=2}^M n_m} & 0\cdots \end{pmatrix}$$

$\in \mathbb{R}^{M-1\times p}$, where each row of the linear combination $\boldsymbol{C\beta}$ corresponds to a Williams-type difference, i.e., a difference of weighted averages of group effects. $n_m, m = 2,\dots M$, refer to the sizes of the treatment groups in ascending dosage order.
Simultaneous confidence intervals for multiple contrasts comparing treatment effects on survival can be obtained by the simultaneous inference procedure by Hothorn et al. (2008) as described in Section 2.4.4. Estimation of the model parameters and the associated covariance matrix will be described in Subsection 4.1.2. Exponentiating the limits of confidence intervals for $\boldsymbol{C\beta}$ yields intervals for hazard ratios or survival time ratios.

When studying effects on survival only a lower bound for survival is of interest. Hence, one-sided confidence intervals are preferred over two-sided intervals (Berger, 2004). Whether upper or lower simultaneous confidence apply depends on the study objective and the choice of survival model. Shorter survival corresponds to a higher effect on the hazard rate in the Cox PH model, but lower effect on time of survival in the AFT model. Hence, if lower confidence limits are used in the Cox PH model, upper confidence intervals apply in the AFT model and vice versa.
For comparison of survival between several experimental groups and a control in clinical trials a negative difference $\beta_m - \beta_1$ of effects from a Cox model corresponds to longer survival for the experimental group $m$ compared to the control group after adjustment for all other explanatory variables in the model. Therefore, upper simultaneous confidence intervals for the Dunnett-type differences need to be used to identify an increase in survival for an experimental group compared to the control.
In toxicological studies, as discussed in the second part of this chapter, the toxic response for the endpoint survival is decrease (equivalent to increase in mortality). For the Cox model a decrease in survival with increasing dosage corresponds to the ordering $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_M, \beta_1 < \beta_M$, for the treatment effects on the hazard rate, with group 1 denoting the control group and groups $2,\dots,M$ denoting the dose groups in ascending order. Thus, lower simultaneous confidence intervals for Williams-type contrasts apply. For the AFT model, a decrease in survival with increasing dosage corresponds to the ordering $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_M, \beta_1 > \beta_M$, for the treatment effects on the survival time requiring upper simultaneous confidence intervals for Williams-type contrasts.

A trend can be established if the interval for at least one contrast does not include 0. The contrast with the corresponding confidence limit most distant from 0 provides information on the dose-response shape.

### 4.1.2 Estimation

In the Cox PH model, estimates of $\boldsymbol{\beta}$ can be obtained by maximization of the partial likelihood function $\mathrm{PL}(\boldsymbol{\beta}; \boldsymbol{x}_i)$ introduced by Cox (1972). The maximum partial likelihood estimates $\hat{\boldsymbol{\beta}}$ are asymptotically multivariate normally distributed

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\beta}))$, the inverse of the expected information matrix. The expectation requires typically nonexistent knowledge about the censoring distribution even for observations with observed events. The expected information matrix can be consistently estimated by the inverse of the observed information matrix $\boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\beta}})$, which will be used to construct simultaneous confidence intervals using the approach by Hothorn et al. (2008) for linear functions of parameters from a Cox PH model in the following.

Parameter estimates in the frailty Cox model are obtained by maximization of a penalized partial log-likelihood function

$$\mathrm{PPL}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{x}_{ij}, \theta) \quad = \quad \mathrm{PL}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{x}_{ij}) - g(\boldsymbol{\gamma}; \theta), \tag{4.1}$$

over both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. A penalty term $g$ constraining the values of the random effects $\boldsymbol{\gamma}$ is subtracted from the Cox partial log-likelihood function PL (for details see Therneau, Grambsch, and Pankratz, 2003). The penalty function depends on the distribution assumed for the frailties. The penalized partial likelihood estimates for the combined vector of fixed and random effects are asymptotically multivariate normal (Parner, 1998). The associated covariance matrix can be consistently estimated by a sandwich matrix $\hat{\boldsymbol{\Sigma}} = \boldsymbol{H}^{-1} \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}) \boldsymbol{H}^{-1}$, with $\boldsymbol{H}^{-1}$ the inverse Hessian matrix of the penalized partial log-likelihood (4.1) and $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})$ the observed Cox information matrix (Gray, 1992).

In the AFT model, estimates of $\boldsymbol{\beta}$ can be obtained by the maximum likelihood method. The maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ are asymptotically normal

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ can be consistently estimated by inverting the observed information matrix obtained from the likelihood function of the AFT model.

## 4.2 Behavior of Dunnett-type Comparisons in the Frailty Cox Model

The following simulation study evaluates the behavior of the simultaneous inference procedure by Hothorn et al. (2008) when applied to multiple comparisons to a control within the frailty Cox model. Therefore, the joint coverage probability of simultaneous confidence intervals is estimated for samples sizes reasonable for phase III studies in settings resembling the data structure of multicenter clinical trials. A coverage of $1 - \alpha$ implies

that the probability of falsely detecting any difference, corresponding to the FWER, is $\alpha$. Bias for maximum (penalized) partial likelihood estimates of (frailty) Cox model parameters in small samples has been reported by Johnson, Tolley, Bryson, and Goldman (1982) and Barker and Henderson (2005). This bias could cause deviation of the joint coverage probability from the nominal confidence level $1-\alpha$ for small samples. We therefore conduct further simulations to examine the small sample properties of maximum penalized partial likelihood estimates of the frailty Cox model.

### 4.2.1 Simulation Setup

All simulations use setups with one control and two experimental groups. According to the considerations in Subsection 4.1.1 one-sided simultaneous confidence intervals, which give an upper bound for the Dunnett-type hazard ratios were used.
Time-to-event data was generated from a Weibull distribution ($\lambda(t|\boldsymbol{x}_i) = \lambda_0(t) \cdot \exp(\boldsymbol{x}_i^\top \beta)$ with $\lambda_0(t) = \lambda\,\nu\,t^{\nu-1}$, $\lambda = 0.5, \nu = 2$) according to Bender, Augustin, and Blettner (2005). For each setting of parameters and sample size, the joint coverage probability of simultaneous confidence intervals for Dunnett-type hazard ratios from a frailty Cox model was estimated based on 10,000 datasets. We considered $\beta_{\text{Exp}_k} - \beta_{\text{Control}}$, $k = 1, 2$, ranging from $-2$ to 0 in steps of 0.1, corresponding to hazard ratios between 0.14 (an 86 % decrease in risk of death) and 1 (no decrease in risk of death). Further covariates were simulated as $x_1 \sim B(1, \frac{1}{2})$ with associated effect $\beta_1 = 0.2$ (corresponding to a sex effect), $x_2$ uniformly distributed on the interval $[18, 65]$ with associated effect $\beta_2 = 0.05$ (corresponding to an age effect), and $x_3 \sim M(1, (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}))$ with $\beta_{31} = -0.3$ and $\beta_{32} = -0.1$ being the effects of categories 1 and 2 compared to the reference category 0. The random effects followed a normal distribution $\gamma_i \sim N(0, 1)$, $i = 1, \ldots, I$. The nominal confidence level was chosen $1 - \alpha = 0.95$.

We considered four designs – one balanced and three unbalanced – with five clusters:

(i) a balanced setting in which all clusters had the same number of observations and the observations were equally distributed to the three treatment groups within each cluster;

(ii) an unbalanced design in which fewer patients were allocated to the control group than to the experimental groups, which can occur for ethical reasons;

(iii) an unbalanced design in which more patients were allocated to the experimental groups;

(iv) an unbalanced design in which the numbers of observations differed between clusters, which can occur, e.g., due to varying recruitment rates in the different study centers.

For each design, total sample sizes of well over 100 and well over 300 were considered, with the exact number of simulated observations depending on the design. Censoring times were simulated from an exponential distribution with the parameter chosen such that censoring rates of approximately 20% were obtained. The allocation of observations

to clusters and treatment groups for the designs (i) to (iv) is shown in Tables 4.1–4.4.

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 7 | 7 | 7 | 7 | 7 |
| Experimental 1 | 7 | 7 | 7 | 7 | 7 |
| Experimental 2 | 7 | 7 | 7 | 7 | 7 |

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 20 | 20 | 20 | 20 | 20 |
| Experimental 1 | 20 | 20 | 20 | 20 | 20 |
| Experimental 2 | 20 | 20 | 20 | 20 | 20 |

Table 4.1: Allocation of simulated observations to treatment groups and clusters in the designs with balanced allocation to clusters and treatments, leading to a total sample size of 105 (left table) and 300 (right table).

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 5 | 5 | 5 | 5 | 5 |
| Experimental 1 | 10 | 10 | 10 | 10 | 10 |
| Experimental 2 | 10 | 10 | 10 | 10 | 10 |

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 15 | 15 | 15 | 15 | 15 |
| Experimental 1 | 30 | 30 | 30 | 30 | 30 |
| Experimental 2 | 30 | 30 | 30 | 30 | 30 |

Table 4.2: Allocation of simulated observations to treatment groups and clusters in the designs with fewer observations in the control group, leading to a total sample size of 125 (left table) and 375 (right table).

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 10 | 10 | 10 | 10 | 10 |
| Experimental 1 | 5 | 5 | 5 | 5 | 5 |
| Experimental 2 | 5 | 5 | 5 | 5 | 5 |

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 30 | 30 | 30 | 30 | 30 |
| Experimental 1 | 15 | 15 | 15 | 15 | 15 |
| Experimental 2 | 15 | 15 | 15 | 15 | 15 |

Table 4.3: Allocation of simulated observations to treatment groups and clusters in the designs with more observations in the control group, leading to a total sample size of 100 (left table) and 300 (right table).

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 3 | 6 | 9 | 12 | 15 |
| Experimental 1 | 3 | 6 | 9 | 12 | 15 |
| Experimental 2 | 3 | 6 | 9 | 12 | 15 |

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Control | 9 | 18 | 27 | 36 | 45 |
| Experimental 1 | 9 | 18 | 27 | 36 | 45 |
| Experimental 2 | 9 | 18 | 27 | 36 | 45 |

Table 4.4: Allocation of simulated observations to treatment groups and clusters in the designs with unbalanced numbers of observations in the five clusters, leading to a total sample size of 135 (left table) and 405 (right table).

Additionally the performance of the simultaneous inference procedure was evaluated in two settings with a high number of clusters and only two observations per cluster with the

observations randomly allocated to one of the three treatments with the same probability for each treatment:

(v) datasets with 120 observations and hence 60 clusters;

(vi) datasets with 300 observations in 150 clusters.

For inspection of potential bias of maximum (penalized) partial likelihood estimates of the (frailty) Cox model parameters in small samples we considered the setups from the simulations on the joint coverage probability. To compare the small sample properties with the asymptotic results, we additionally considered sample sizes of approximately 10,000 in the balanced and unbalanced designs. The asymptotic variance of the penalized partial likelihood estimates was obtained the same way as in Johnson et al. (1982): The 10,000 samples, in each of which the log hazard ratios and their asymptotic variances had been estimated, were split into 10 groups each containing 1,000 samples. In each group the median of the 1,000 asymptotic variances of each sample was calculated. The asymptotic variance was estimated by the average of the 10 medians.

### 4.2.2 Simulation Results

#### Joint Coverage Probability

In the four designs with five clusters (i)–(iv), the joint coverage probability of the simultaneous confidence intervals for the hazard ratios comparing both experimental treatments with the standard treatment decreases with decreasing hazard ratios for both the settings with smaller and with larger sample size. The estimated coverage probability is illustrated in Figures 4.1–4.4. For sample sizes of approx. 100, the joint coverage probability ranges between 0.91 and 0.94 for most of the parameter combinations and is thus considerably below the nominal level of 0.95. For sample sizes of approx. 300, the estimated coverage probability is close to the nominal 0.95.

There was no systematic pattern of the estimated coverage probability and the values of the treatment effects in the settings with higher number of groups (v)-(vi). For 120 observations in 60 clusters, the coverage probability ranges from 0.93 to 0.939; for sample sizes of 300 in 150 clusters, the joint coverage probability ranges from 0.948 to 0.963.

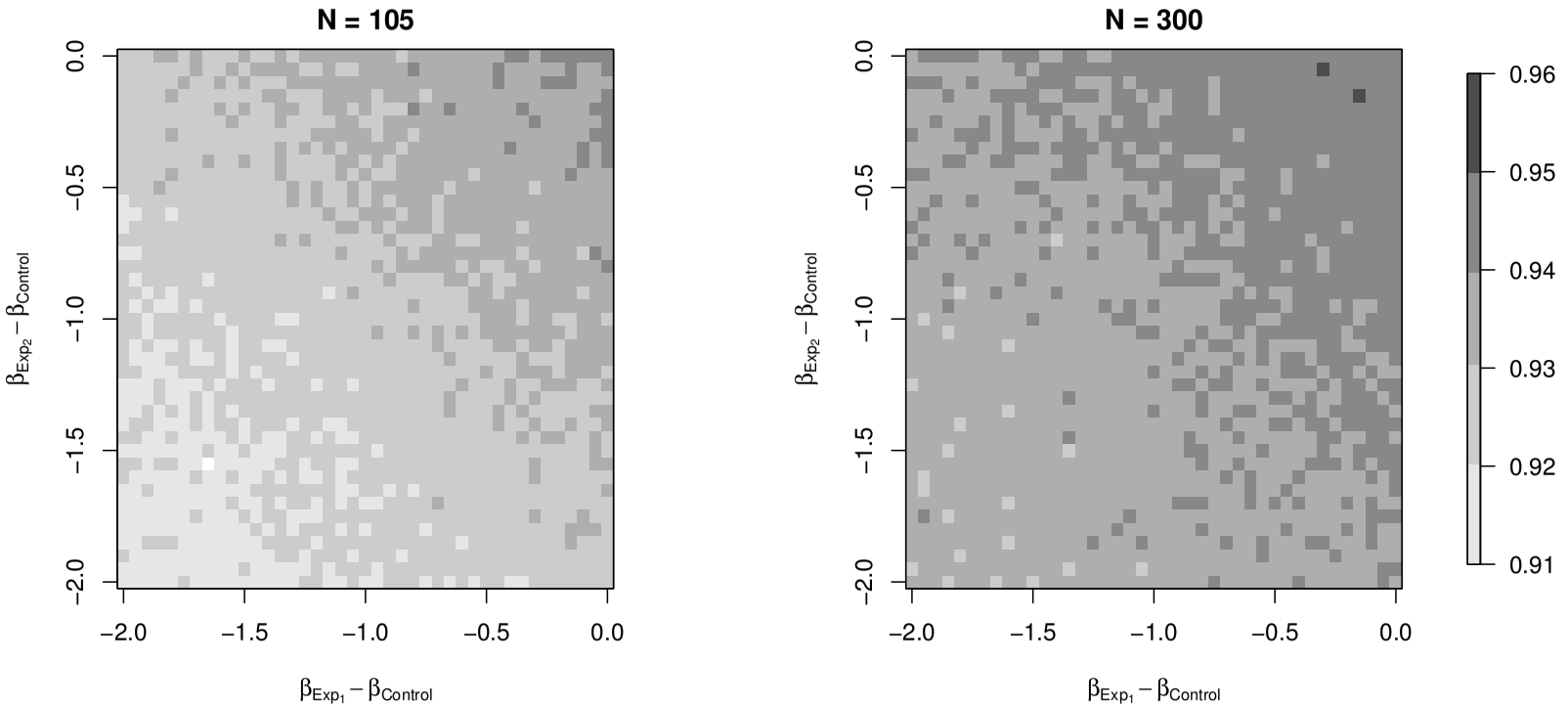


Figure 4.1: Estimated joint coverage probability of 95% upper simultaneous confidence intervals for Dunnett-type contrasts comparing each of two experimental treatments with a control in the frailty Cox model for a balanced design.

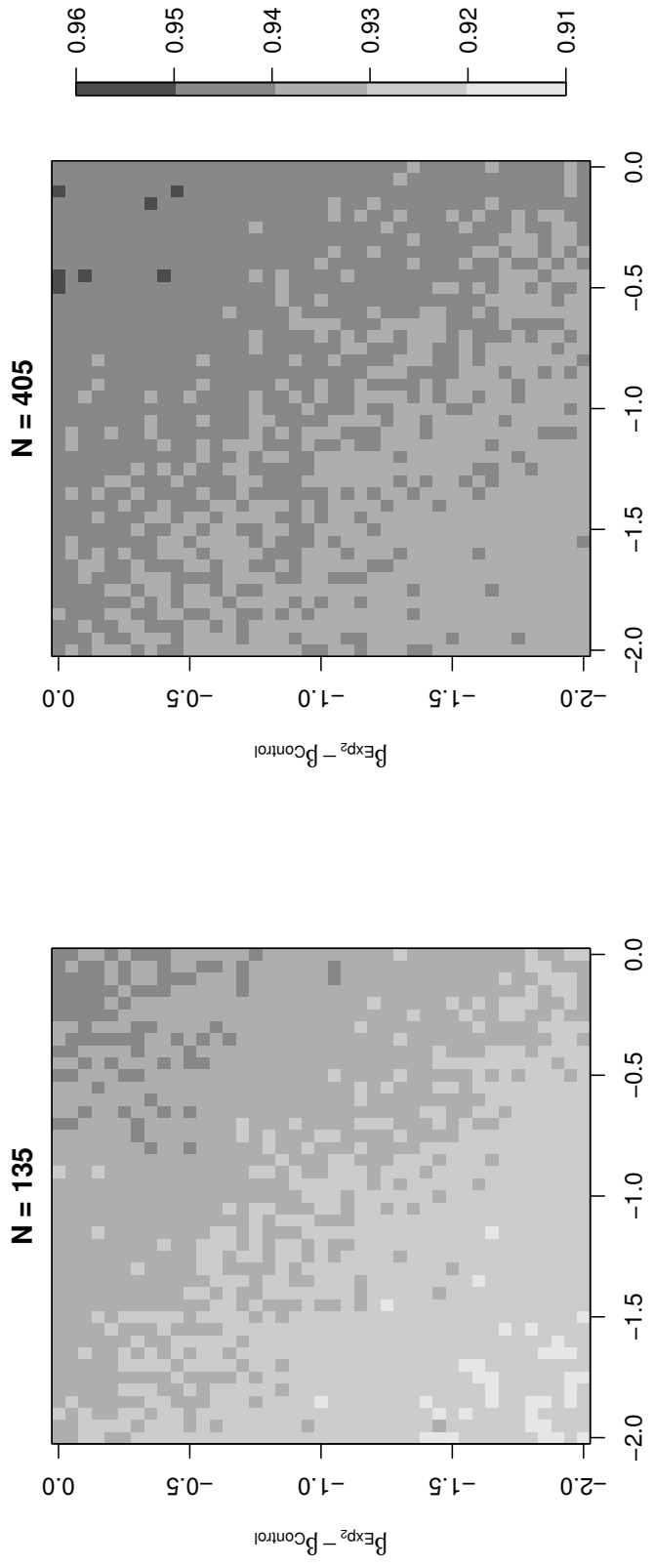


Figure 4.2: Estimated joint coverage probability of 95% upper simultaneous confidence intervals for Dunnett-type contrasts comparing each of two experimental treatments with a control in the frailty Cox model for an unbalanced design with fewer patients allocated to the control group.

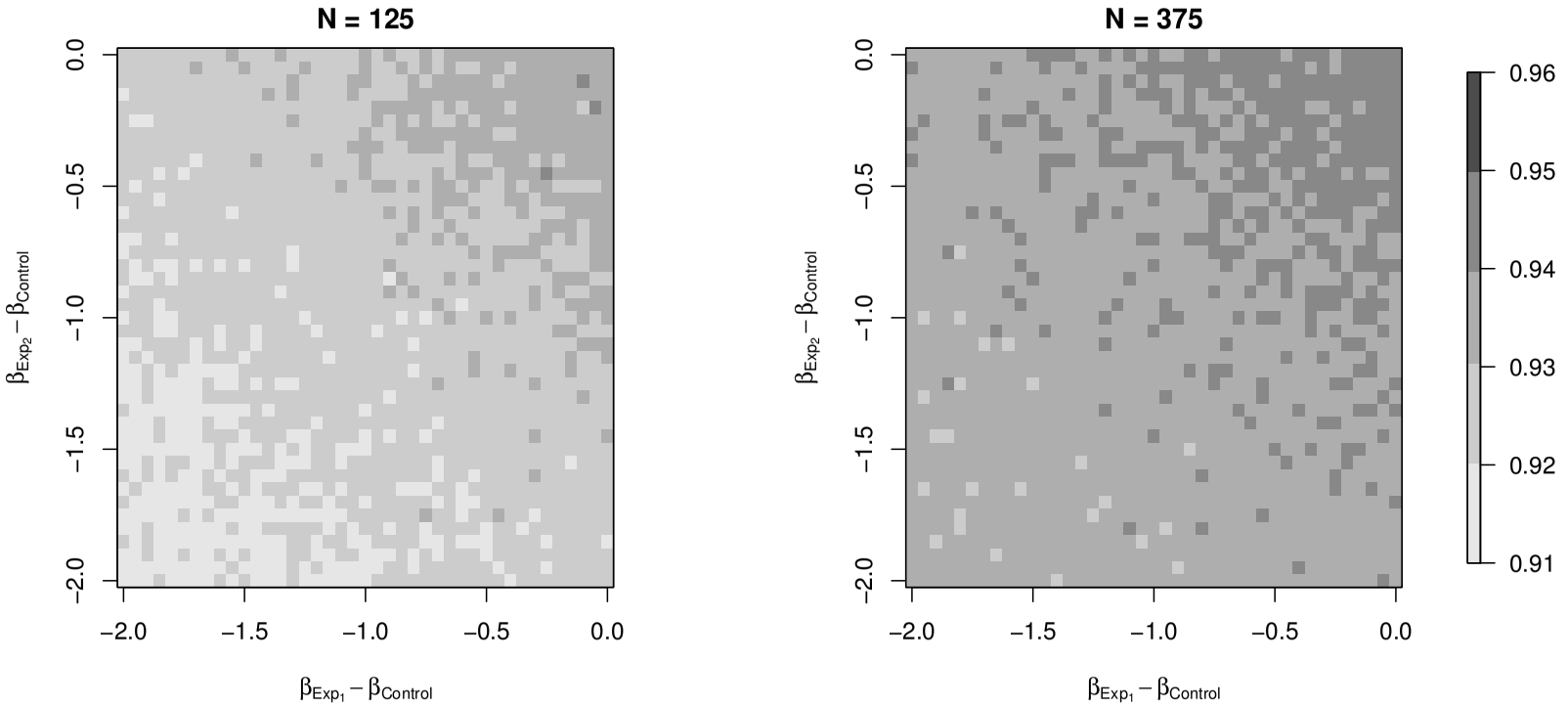


Figure 4.3: Estimated joint coverage probability of 95% upper simultaneous confidence intervals for Dunnett-type contrasts comparing each of two experimental treatments with a control in the frailty Cox model for an unbalanced design with more patients allocated to the control group.

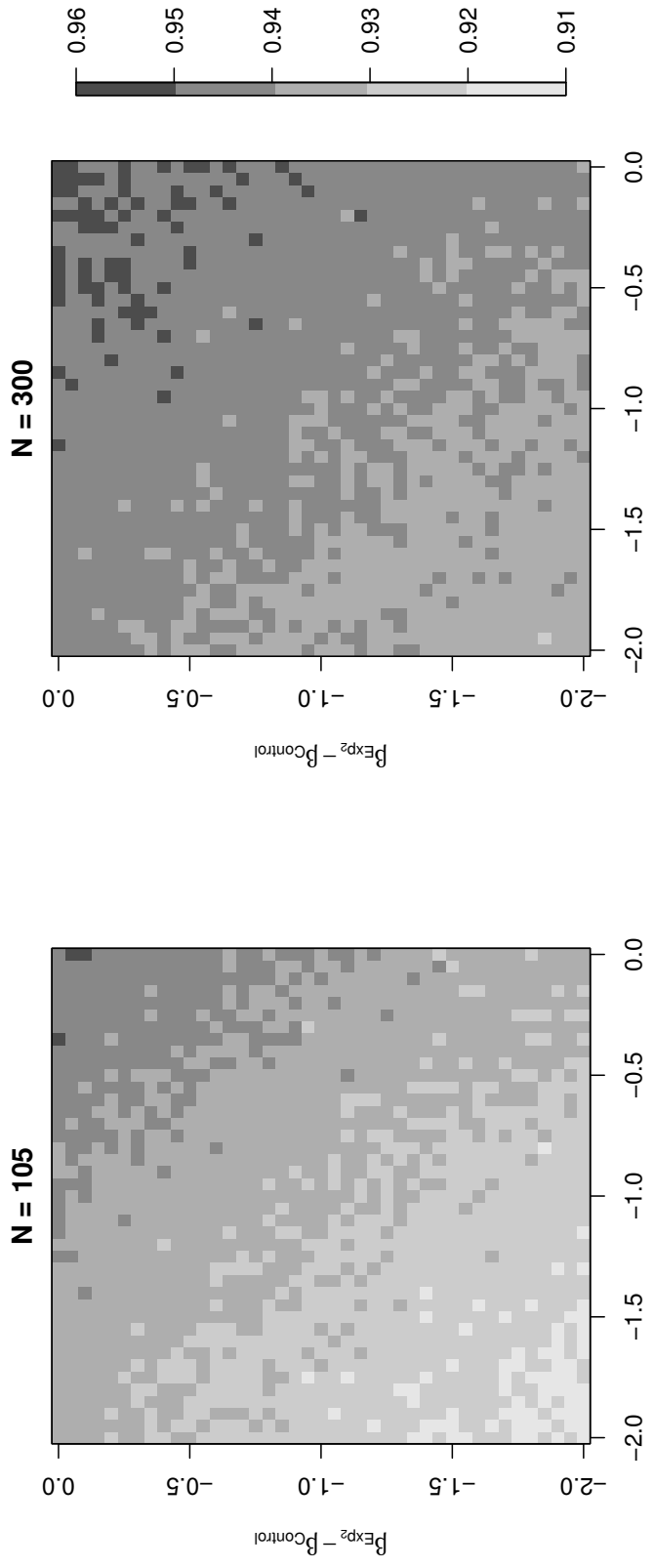


Figure 4.4: Estimated joint coverage probability of 95% upper simultaneous confidence intervals for Dunnett-type contrasts comparing each of two experimental treatments with a control in the frailty Cox model for an unbalanced design with the number of observations differing between clusters.

### Bias of Parameter Estimates

The joint coverage probability of the simultaneous confidence intervals decreases systematically with increasing distance of the effect size from the null hypothesis. Undercoverage of the simultaneous confidence intervals can result from bias of the parameter estimates. The distribution of $\hat{\beta}_{\text{Exp}_1}$ for true values $\beta_{\text{Exp}_1} = -2, -1, 0, 2$ and $\beta_{\text{Exp}_1} = \beta_{\text{Exp}_2}$ is shown in Figures 4.5 (balanced setting), 4.6 (unbalanced setting with less observations in the control), and 4.7 (unbalanced setting with more observations in the control). The blue curves show the estimated densities of $\hat{\beta}_{\text{Exp}_1}$ obtained by kernel density estimation using the 10,000 estimated values. The black curves show the normal density curves with mean $\beta_{\text{Exp}_1}$ and estimated asymptotic variance of $\hat{\beta}_{\text{Exp}_1}$.

For sample sizes of approximately 100 and 300 the shape of the kernel estimated density curves agrees closely with the shape of the "true" normal curves, but the estimated curves are shifted for $\beta_{\text{Exp}_1} \neq 0$. The bias of the estimates increases with increasing distance of the true log hazard ratio values from zero. Estimated log hazard ratios are unbiased in samples with approx. 10,000 observations, and the bias reduces considerably when the sample size is increased from 100 to 300 or higher. The bias is slightly larger in the balanced setting compared to the unbalanced settings, which might be due to the smaller number of observations (100 vs. 125 or 135 and 300 vs. 375 or 405).

Table 4.5 shows the bias of the estimated log hazard ratio $\hat{\beta}_1$, the finite sample variance and the asymptotic variance of the estimated log hazard ratios. The finite sample variance is slightly larger than the asymptotic variance in all considered settings. These results confirm the findings by Johnson et al. (1982) and agree with the undercoverage of the simultaneous confidence intervals.

Randomized trials in which hazard ratios are used as effect size usually report hazard ratios between 0.5 and 2 or respectively log hazard ratios between $-0.7$ and 0.7 (see for example Saltz et al., 2001; Moore et al., 2007; Butts et al., 2010). In this range of log hazard ratios, the bias of the estimates is relatively small and the estimated coverage of the confidence intervals is very close to $1 - \alpha$ for sample sizes of approximately $N = 300$.

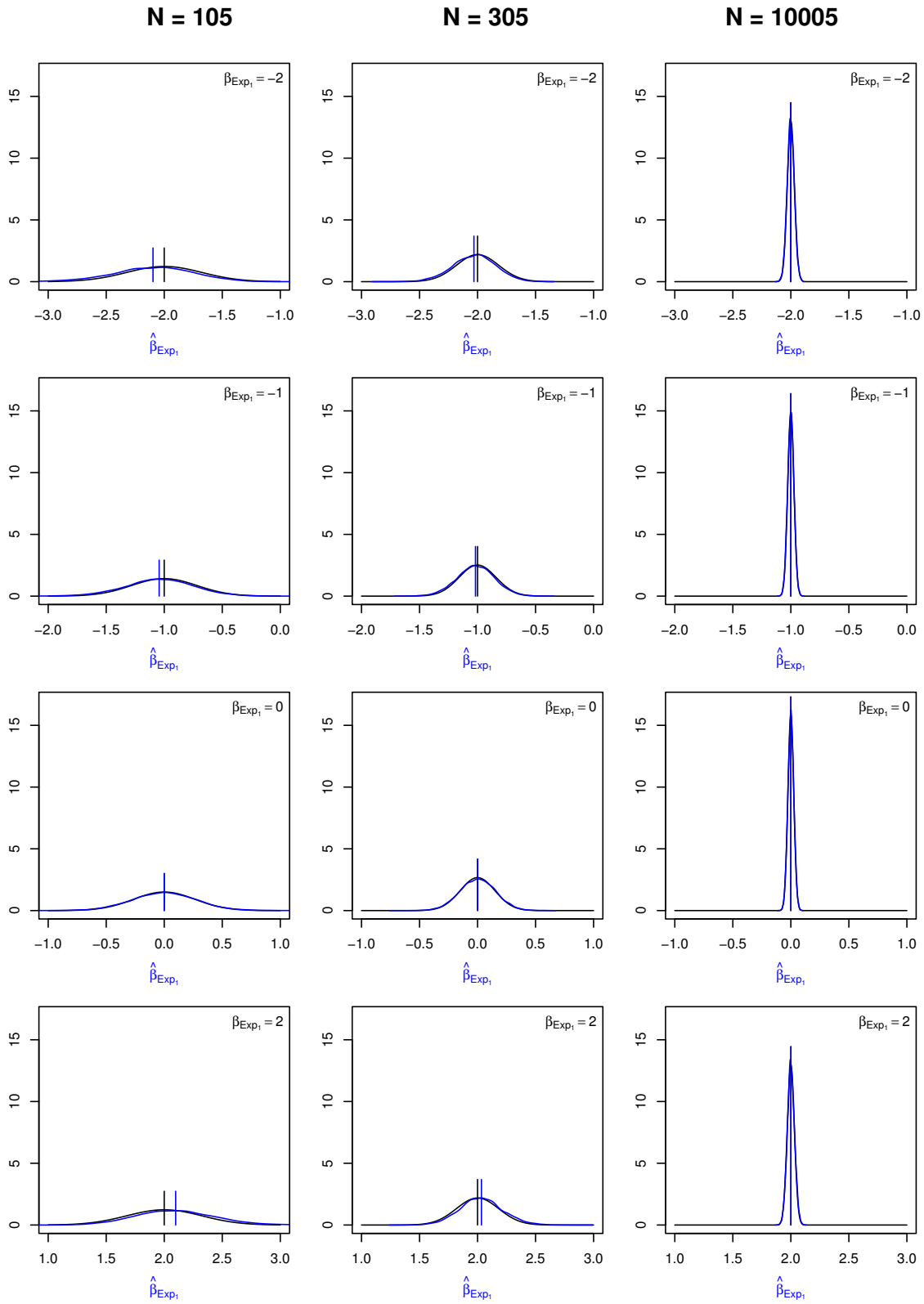


Figure 4.5: Distribution of $\hat{\beta}_{\text{Exp}_1}$ for true values $\beta_{\text{Exp}_1} = -2, -1, 0, 2$ and $\beta_{\text{Exp}_1} = \beta_{\text{Exp}_2}$ in the balanced setting with total sample sizes of $N = 105$, 300, and 10005.

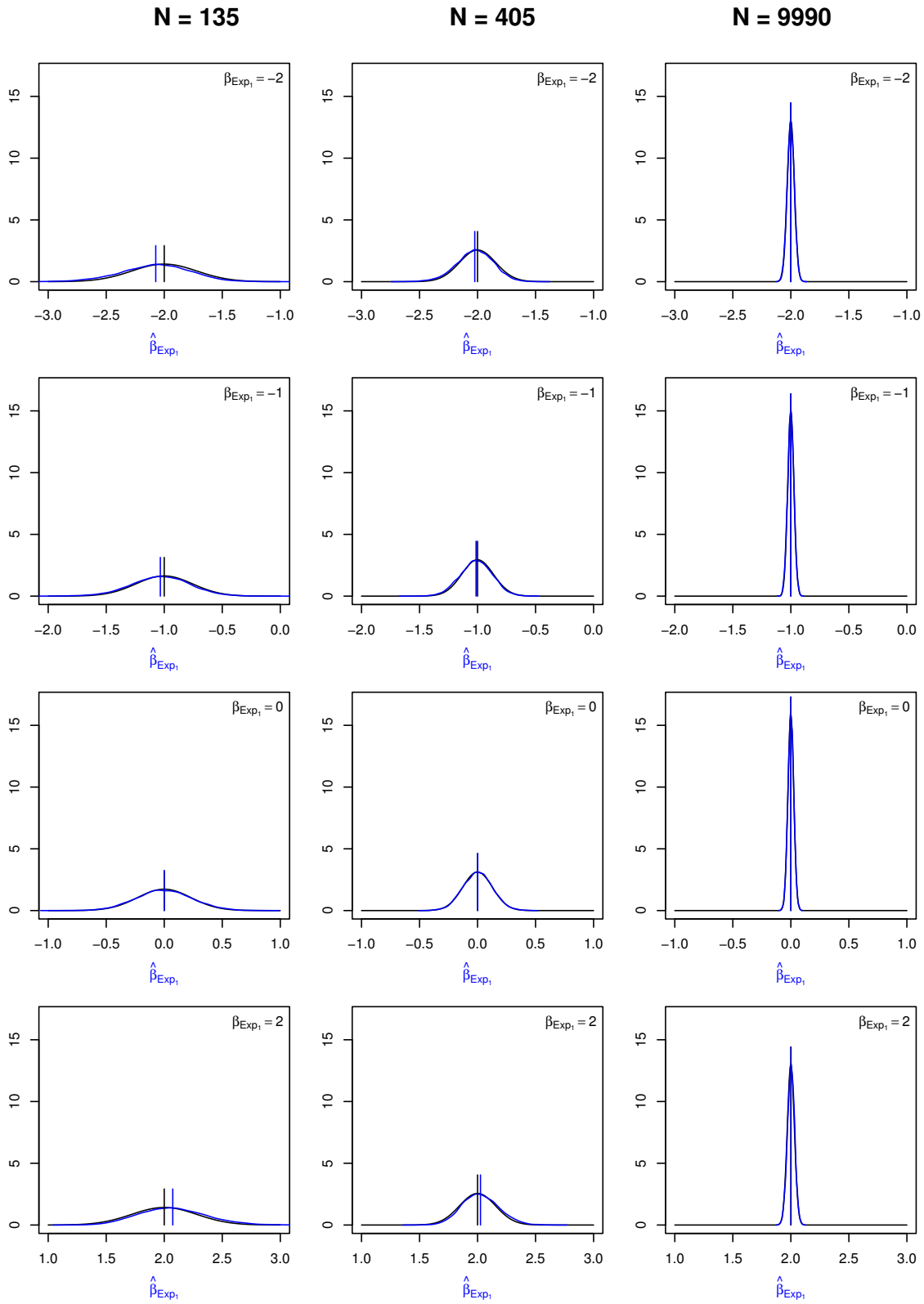


Figure 4.6: Distribution of $\hat{\beta}_{\text{Exp}_1}$ for true values $\beta_{\text{Exp}_1} = -2, -1, 0, 2$ and $\beta_{\text{Exp}_1} = \beta_{\text{Exp}_2}$ in the unbalanced setting with less observations in the control with total sample sizes of $N = 135$, 405, and 9990.

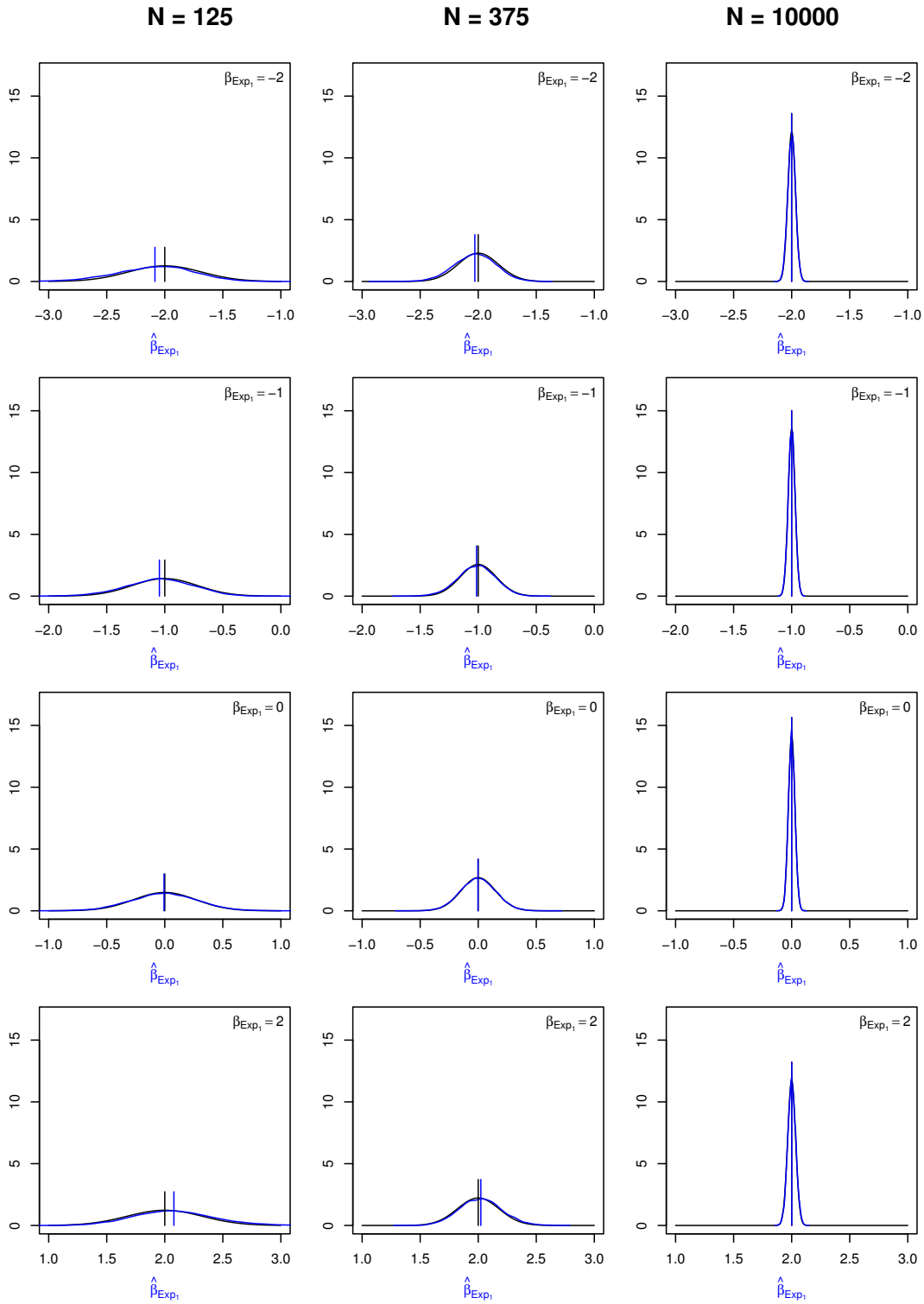


Figure 4.7: Distribution of $\hat{\beta}_{\text{Exp}_1}$ for true values $\beta_{\text{Exp}_1} = -2, -1, 0, 2$ and $\beta_{\text{Exp}_1} = \beta_{\text{Exp}_2}$ in the unbalanced setting with more observations in the control with total sample sizes of $N = 125$, 375, and 10000.

| | $\beta_1$ | $\beta_2$ | $N = 105$ bias($\hat{\beta}_1$) | $N = 105$ finite-sample variance | $N = 105$ asymptotic variance | $N = 300$ bias($\hat{\beta}_1$) | $N = 300$ finite-sample variance | $N = 300$ asymptotic variance |
|---|---|---|---|---|---|---|---|---|
| balanced | -2 | -2 | -0.097 | 0.121 | 0.105 | -0.031 | 0.034 | 0.033 |
| | -1 | -2 | -0.043 | 0.088 | 0.079 | -0.016 | 0.026 | 0.025 |
| | 0 | -2 | -0.003 | 0.079 | 0.070 | -0.000 | 0.023 | 0.022 |
| | -1 | -1 | -0.044 | 0.088 | 0.079 | -0.018 | 0.026 | 0.025 |
| | 0 | 0 | -0.000 | 0.077 | 0.071 | 0.002 | 0.024 | 0.022 |
| | 1 | 1 | 0.053 | 0.088 | 0.078 | 0.018 | 0.026 | 0.025 |
| | 2 | 2 | 0.099 | 0.118 | 0.105 | 0.034 | 0.035 | 0.033 |
| less observations | -2 | -2 | -0.073 | 0.088 | 0.079 | -0.024 | 0.025 | 0.024 |
| in control | -1 | -2 | -0.035 | 0.065 | 0.059 | -0.008 | 0.019 | 0.018 |
| | 0 | -2 | 0.001 | 0.056 | 0.053 | 0.001 | 0.017 | 0.016 |
| | -1 | -1 | -0.034 | 0.064 | 0.060 | -0.013 | 0.019 | 0.018 |
| | 0 | 0 | -0.000 | 0.058 | 0.053 | 0.002 | 0.016 | 0.016 |
| | 1 | 1 | 0.038 | 0.064 | 0.059 | 0.013 | 0.019 | 0.018 |
| | 2 | 2 | 0.073 | 0.084 | 0.079 | 0.025 | 0.025 | 0.024 |
| more observations | -2 | -2 | -0.085 | 0.111 | 0.100 | -0.029 | 0.032 | 0.030 |
| in control | -1 | -2 | -0.047 | 0.087 | 0.079 | -0.014 | 0.025 | 0.024 |
| | 0 | -2 | -0.015 | 0.078 | 0.072 | -0.004 | 0.023 | 0.022 |
| | -1 | -1 | -0.046 | 0.085 | 0.079 | -0.015 | 0.026 | 0.024 |
| | 0 | 0 | -0.006 | 0.082 | 0.072 | -0.002 | 0.023 | 0.022 |
| | 1 | 1 | 0.031 | 0.088 | 0.080 | 0.013 | 0.025 | 0.024 |
| | 2 | 2 | 0.079 | 0.118 | 0.105 | 0.023 | 0.033 | 0.032 |

Table 4.5: Estimated bias, finite-sample variance and asymptotic variance for $\hat{\beta}_1$.

## 4.3 Comparisons of Therapies for Chronic Myelogenous Leukemia

Chemotherapy with the cytostatic drugs busulfan (BUS) or hydroxyurea (HU) had long been the standard treatments for chronic myelogenous leukemia (CML) before the cytokine Interferon-$\alpha$ (IFN-$\alpha$) was introduced to the CML therapy (Bolin, Robinson, Sutherland, and Hamman, 1982). The German CML Study Group compared the patient survival of IFN-$\alpha$ with that of the two conventional chemotherapies and other secondary endpoints in a randomized trial (Hehlmann et al., 1994). From 1983 to 1991, a total of 516 eligible patients were recruited in 57 study centers and randomized to one of the three treatments. Of 516, 507 had data on sex, age and a prognostic score distinguishing three risk groups (low, intermediate, and high) (Hasford et al., 1998). These 507 patients had been randomly assigned to the treatments as follows: 132 to IFN-$\alpha$, 182 to BUS, and 193 to HU. In an earlier analysis in 1994, log-rank tests with Bonferroni correction for multiplicity indicated a significant difference in survival between IFN-$\alpha$ and BUS, but no significant difference between IFN-$\alpha$ and HU (Hehlmann et al., 1994). We reanalyzed the data and compared IFN-$\alpha$ with both BUS and HU using simultaneous confidence intervals for the corresponding hazard ratios in the frailty Cox model with adjustment for covariates and potential heterogeneity between the centers. In the original analysis 148 patients were still under risk. The observations of these patients were updated for the new analysis. Of the 507 patients, 417 died; the observations of the remaining 90 patients are censored, mainly due to bone marrow transplantation in the first chronic phase. We fitted the data using a frailty Cox model with Gaussian frailties age for the centers and included the covariates treatment, sex, age, and a prognostic score. In the present situation, where two standard treatments are compared with one experimental treatment, hazard ratios $\exp(\beta_{\text{Stand}} - \beta_{\text{Exp}})$ greater than 1 correspond to a lower risk of death for the experimental treatment compared to the standard. Hence, one-sided lower simultaneous confidence intervals giving a lower bound for the hazard ratios were computed. The hazard ratios were estimated as $\exp(\beta_{\text{BUS}} - \beta_{\text{INF}\alpha}) = 1.52$ (52% higher risk of death for BUS compared to IFN-$\alpha$), and $\exp(\beta_{\text{HU}} - \beta_{\text{INF}\alpha}) = 1.34$ (34% higher risk of death for HU compared to IFN-$\alpha$). The 95% lower simultaneous confidence intervals resulted in $[1.19, \infty)$ for the hazard ratio of BUS compared to IFN-$\alpha$ and $[1.05, \infty)$ for the hazard ratio of HU compared to IFN-$\alpha$, which indicated significant differences between the hazards of both standard treatments and IFN-$\alpha$, even though the decrease in risk of death with IFN-$\alpha$ compared to HU might not be of clinical importance.

## 4.4 Behavior of Williams-type Comparisons in the Cox PH and AFT Model

The following simulation study aims to assess the behavior of the simultaneous inference procedure by Hothorn et al. (2008) when applied to trend analyses within the Cox PH model and the AFT model using simultaneous confidence intervals for Williams-type con-

trasts. Scenarios resembling the data structure of toxicological studies were used. First, the joint coverage probability of the intervals under the null hypothesis, i.e. no difference between treatment effects was inspected. A coverage of $1-\alpha$ implies that the probability of falsely detecting a trend of any form, corresponding to the FWER, is $\alpha$.
Power analyses are divided in two parts. The first part investigates how the power properties are

- in balanced settings compared to unbalanced settings with either more observations in the control group or less observations in the control group compared to the dosage groups,
- for settings with a higher event rate compared to a lower event rate.

The second part investigates the power of simultaneous confidence intervals for Williams-type contrasts

- compared to Dunnett-type contrasts, which represent many-to-one comparisons with the control,
- for three different dose-response shapes (convex, linear, concave).

### 4.4.1 Simulation Setup

Sample sizes in toxicological studies are usually rather small, with approximately 50 observations in the dose groups and the same number or fewer observations in the control group. We considered balanced and unbalanced designs with sample sizes as given in Table 4.6 for estimation of the joint coverage probability under the null hypothesis.

| Design | $n_1$ | $n_2$ | $n_3$ | $n_4$ |
|---|---|---|---|---|
| Balanced 1 | 30 | 30 | 30 | 30 |
| Balanced 2 | 50 | 50 | 50 | 50 |
| Less in Control 1 | 12 | 30 | 30 | 30 |
| Less in Control 2 | 20 | 50 | 50 | 50 |
| More in Control 1 | 42 | 30 | 30 | 30 |
| More in Control 2 | 70 | 50 | 50 | 50 |

Table 4.6: Allocation of the observations to $M = 4$ groups in the simulations.

Mortality should be analyzed by one-sided tests or confidence intervals for a possible increase, i.e., lower simultaneous confidence intervals for Williams-type contrasts in the Cox PH model or upper simultaneous confidence intervals in the AFT model. The current approach of the National Toxicology Programm for analysis of mortality in long-term carcinogenicity bioassays is to compute a two-sided $p$-value for Tarone's global test of a linear trend. To stay compatible with the NTP practice in terms of the level we use one-sided simultaneous confidence intervals with nominal confidence level $1-\alpha/2 = 0.975$.
Throughout the simulations $M = 4$ groups were compared, leading to three Williams-type contrasts. Time-to-event data were generated from a Weibull distribution ($\lambda(t) =$

$\lambda_0(t) \cdot \exp(\beta_i)$ with $\lambda_0(t) = \lambda \nu t^{\nu-1}$) according to Bender et al. (2005), with scale parameter $\lambda = 1/1,000$ throughout the simulations. Two different shape parameters were used, leading to fewer events ($\nu = 1$) or more events ($\nu = 1.2$). Long-term carcinogenicity bioassays run for approximately two years. The times of all animals alive until then are censored. In the simulations fixed censoring was applied at time $t = 700$. Five observations were censored early. These censoring times were uniformly distributed over $[0, 700]$ and randomly allocated to the groups. We did not include any further covariates, since possible covariates such as body-weight vary over time.
To assess the joint coverage probability of the Williams-type simultaneous confidence intervals, data were generated under the null hypothesis $\beta_1 = \ldots = \beta_4 = 0$. For Williams-type contrasts from the Cox PH model, the joint coverage probability was estimated by the portion among 10,000 simulated data sets in which at least one of the corresponding lower simultaneous confidence intervals excludes the value 0. For Williams-type contrasts from the AFT model assuming Weibull distributed times, the joint coverage probability was estimated by the portion among 10,000 simulated data sets in which at least one of the corresponding upper simultaneous confidence intervals excludes the value 0.

For the power analyses datasets were simulated with a trend present. In the first part of the power analyses data were simulated for the same settings and parameter values like before but with a trend present following a convex dose-response shape $\beta_1 = \beta_2 = \beta_3 < \beta_4$. The considered effects were $\beta = (-0.01 \cdot l, -0.01 \cdot l, -0.01 \cdot l, 0.03 \cdot l)$ with $l$ ranging from 0.5 to 20 in steps of 0.5, leading to parameter vectors from $\beta = (-0.005, -0.005, -0.005, 0.015)$ up to $\beta = (-0.2, -0.2, -0.2, 0.6)$. The corresponding hazard ratios comparing the highest dose group with the control group range from $\exp(\beta_4 - \beta_1) = 1.02$ to $\exp(\beta_4 - \beta_1) = 2.23$. The power was estimated by the portion of data sets where at least one of the simultaneous intervals for Williams-type contrasts does not include 0, i.e. a trend is detected.

In the second part of the power analyses only balanced designs with either 50 or 30 objects in each group were used. In addition to simultaneous confidence intervals for Williams-type contrasts, simultaneous confidence intervals for Dunnett-type contrasts were computed to compare the power to detect a trend between these two approaches. Convex dose-response shapes like in the first part of the power simulations were used and additionally linear shapes $\beta_1 < \beta_2 < \beta_3 < \beta_4$ and concave shapes $\beta_1 < \beta_2 = \beta_3 = \beta_4$ were inspected. For the linear shapes the effects were chosen $\beta = (-0.02 \cdot l, -0.0067 \cdot l, 0.0067 \cdot l, 0.02 \cdot l)$ with $l$ ranging from 0.5 to 20 in steps of 0.5, leading to parameter vectors from $\beta = (-0.01, -0.0033, 0.0033, 0.01)$ up to $\beta = (-0.4, -0.13, 0.13, 0.4)$. For the concave shapes the effects were chosen $\beta = (-0.03 \cdot l, 0.01 \cdot l, 0.01 \cdot l, 0.01 \cdot l)$ with $l$ ranging from 1 to 20 in steps of 0.5, leading to parameter vectors from $\beta = (-0.015, 0.005, 0.005, 0.005)$ up to $\beta = (-0.6, 0.2, 0.2, 0.2)$. The corresponding hazard ratios comparing the highest dose group with the control group ranged from $\exp(\beta_4 - \beta_1) = 1.02$ to $\exp(\beta_4 - \beta_1) = 2.23$ for all three shapes.

### 4.4.2 Simulation Results

#### Joint Coverage Probability

The estimated joint coverage probability under the null hypothesis for the different designs is shown in Figure 4.8. The coverage decreases with increasing event rate ($\nu = 1.2$ compared to $\nu = 1$). The simultaneous confidence intervals are slightly conservative for fewer observations in the control group, whereas the joint coverage probability reaches the nominal confidence level for the balanced design and the design with more observations in the control. Coverage of the intervals for Williams-type contrasts from the AFT model is slightly higher than the coverage of the corresponding intervals from the Cox PH model for $\nu = 1.2$, and slightly lower for $\nu = 1.0$.

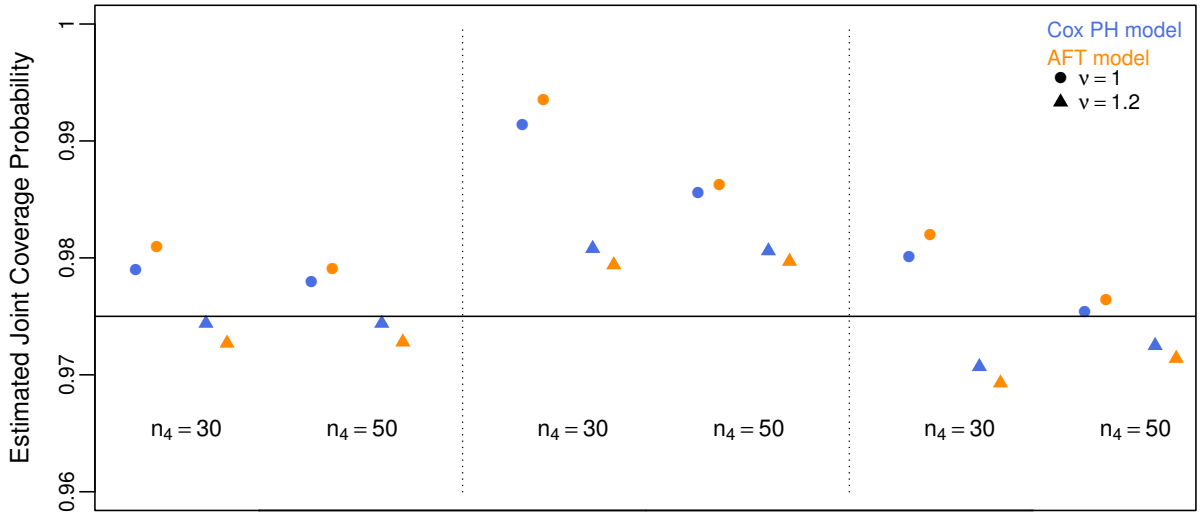


Figure 4.8: Estimated joint coverage probability of 97.5% lower simultaneous confidence intervals for the Williams-type contrasts.

#### Power

The power curves for the settings of the first part of the power analyses are shown in Figures 4.9 and 4.10. The results of the simulated power calculation agree with the findings of the coverage probability estimations under the null hypothesis. In the balanced setting, which is recommended by the NTP, sample sizes of 30 or 50 per group yield a good power to detect a trend when the dose-response shape is convex. The power of the procedure is the lowest for designs with fewer observations in the control group and highest for designs with more observations in the control group. However, for designs with more observations in the control group, the nominal coverage probability was not reached by data simulated from a Weibull distribution with shape parameter $\nu = 1.2$. A considerable increase in power can be achieved when increasing the sample size of the highest dose group from $n_4 = 30$ to $n_4 = 50$ and the sample sizes of the other groups according to the designs (Figure 4.10 versus 4.9). There is no essential difference between the power when modeling survival by an AFT model compared to a Cox PH model.

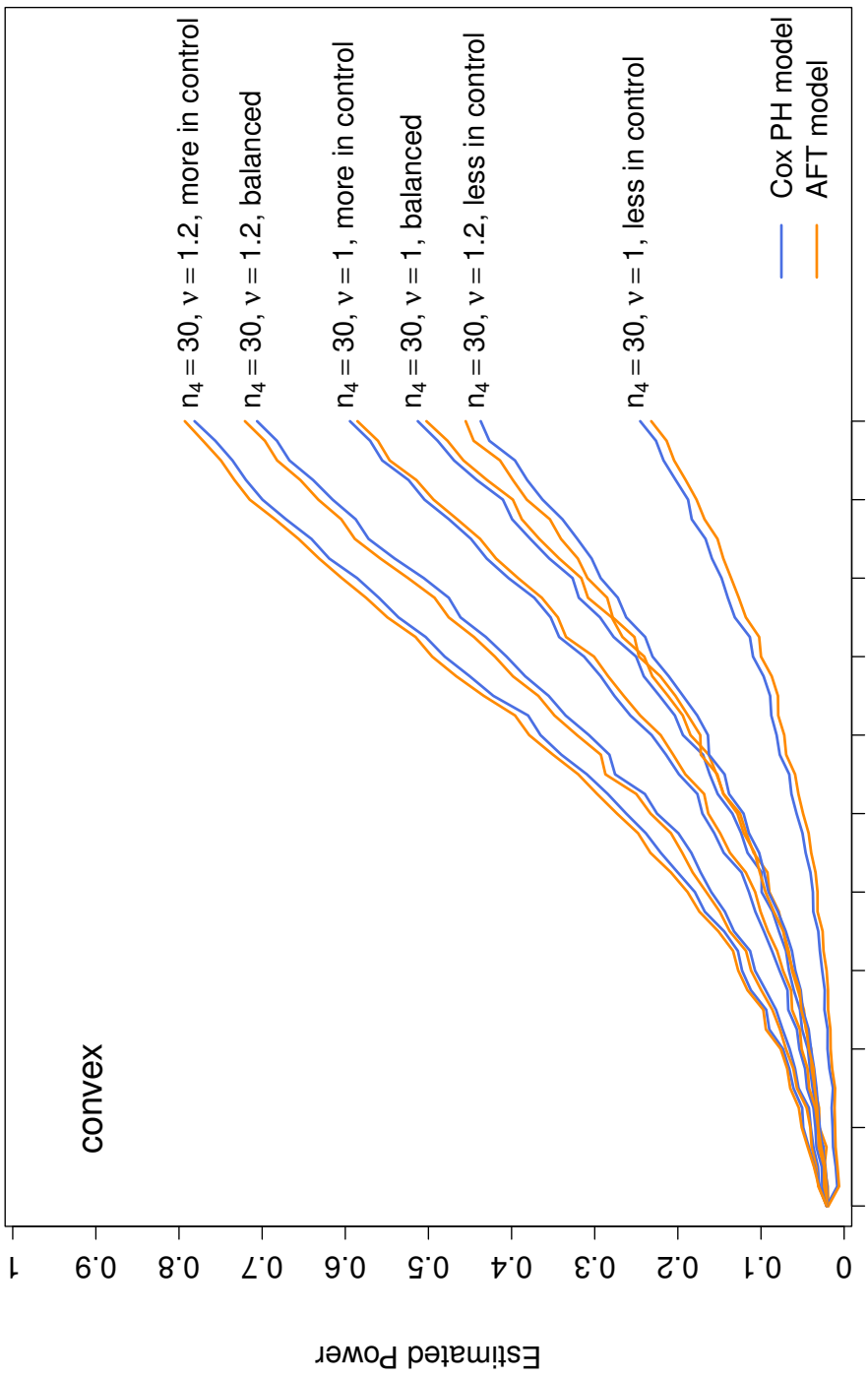


Figure 4.9: Estimated power of the Williams-type procedure using one-sided 97.5% simultaneous confidence intervals in data sets with $n_4 = 30$ objects in dosage group 4.

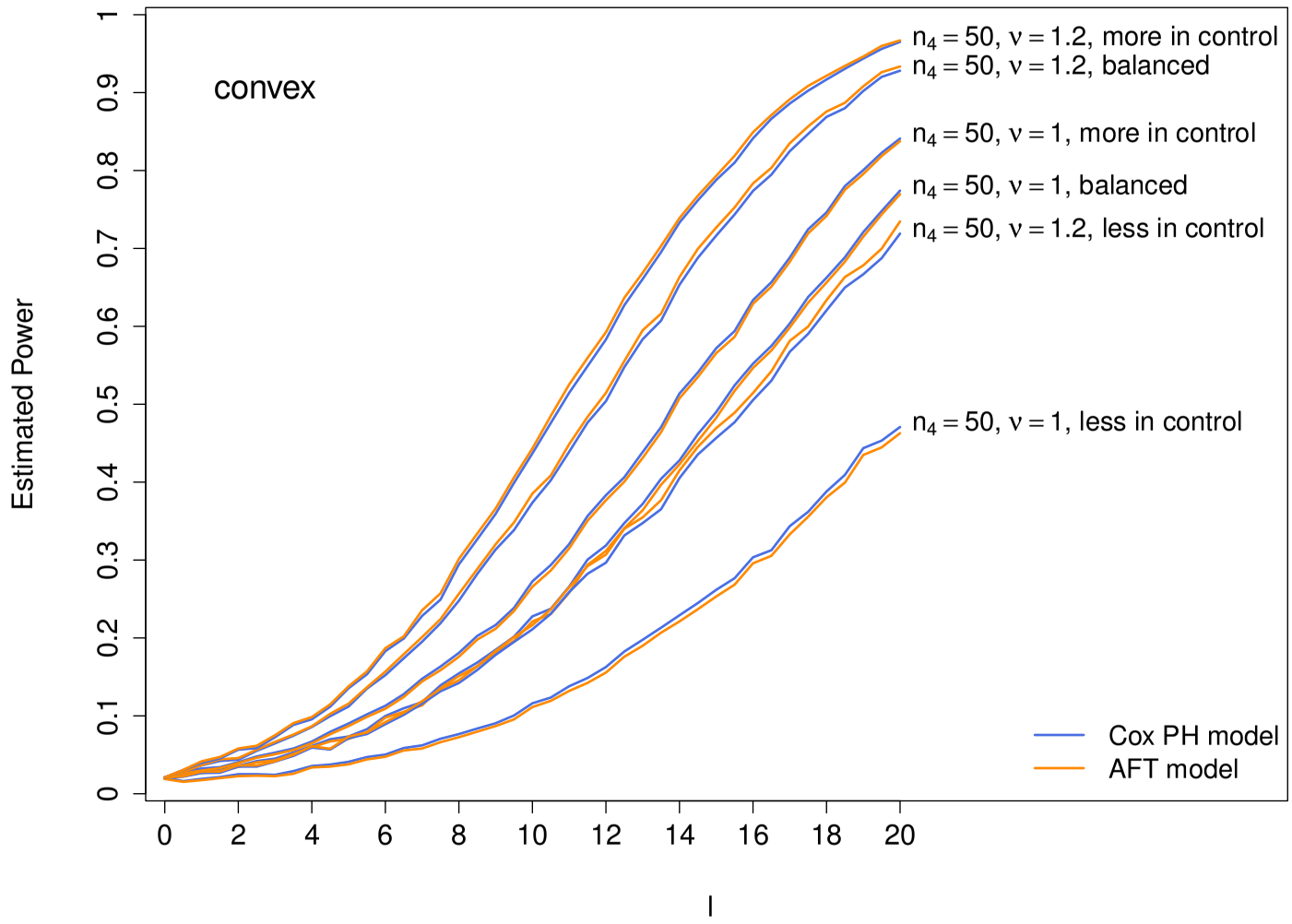


Figure 4.10: Estimated power of the Williams-type procedure using one-sided 97.5% simultaneous confidence intervals in data sets with $n_4 = 50$ objects in dosage group 4.

The power curves for the settings of the second part of the power analyses are shown in Figures 4.11 and 4.12.

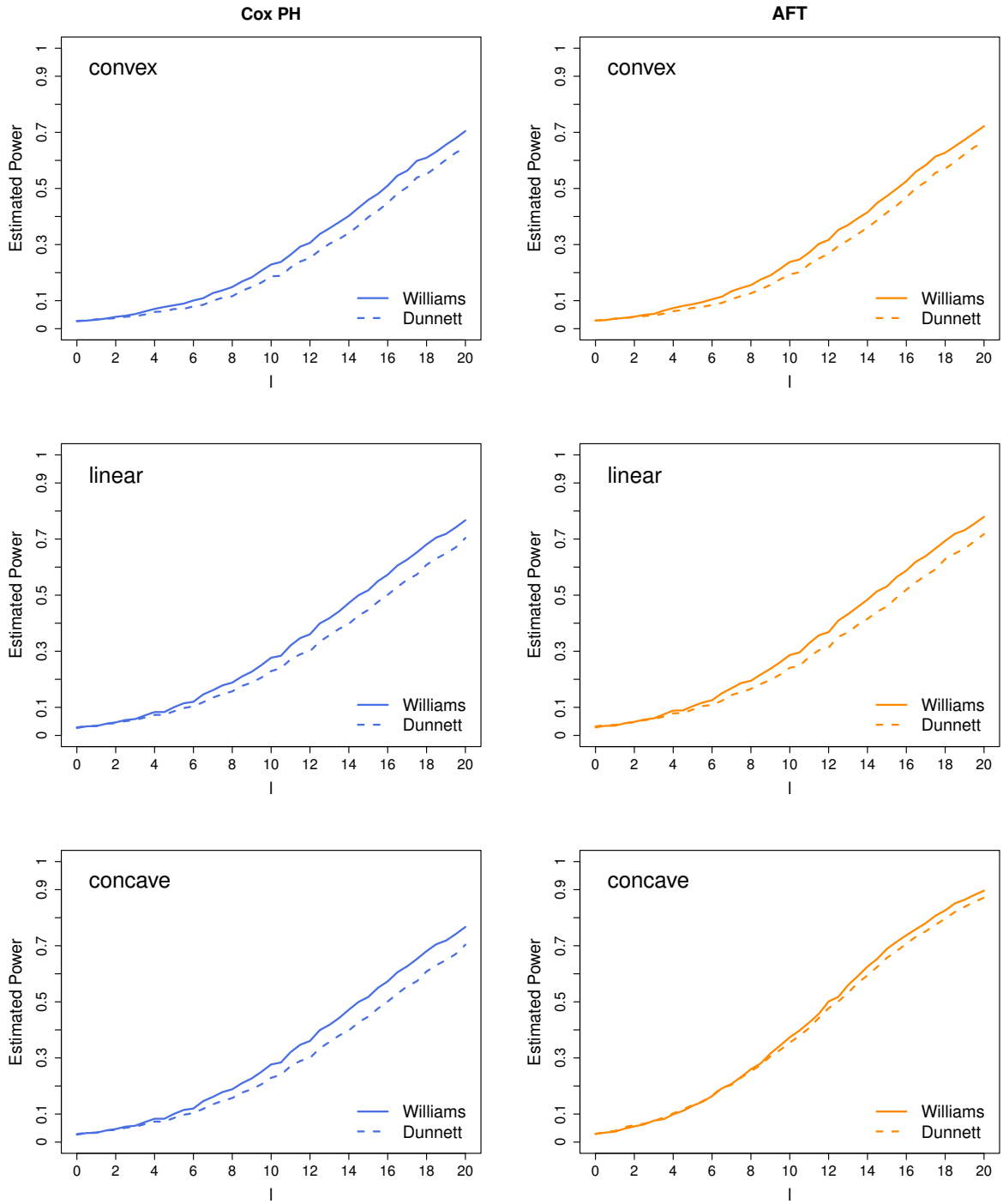


Figure 4.11: Estimated power of the one-sided 97.5% simultaneous confidence intervals for Williams-type contrasts versus Dunnett-type contrasts for a convex dose-response shape (upper row), linear dose-response shape (middle row), and concave dose-response shape (bottom row) for the Cox PH model (left column) and the AFT model (right column) in balanced settings with $n_4 = 30$ objects in dosage group 4.

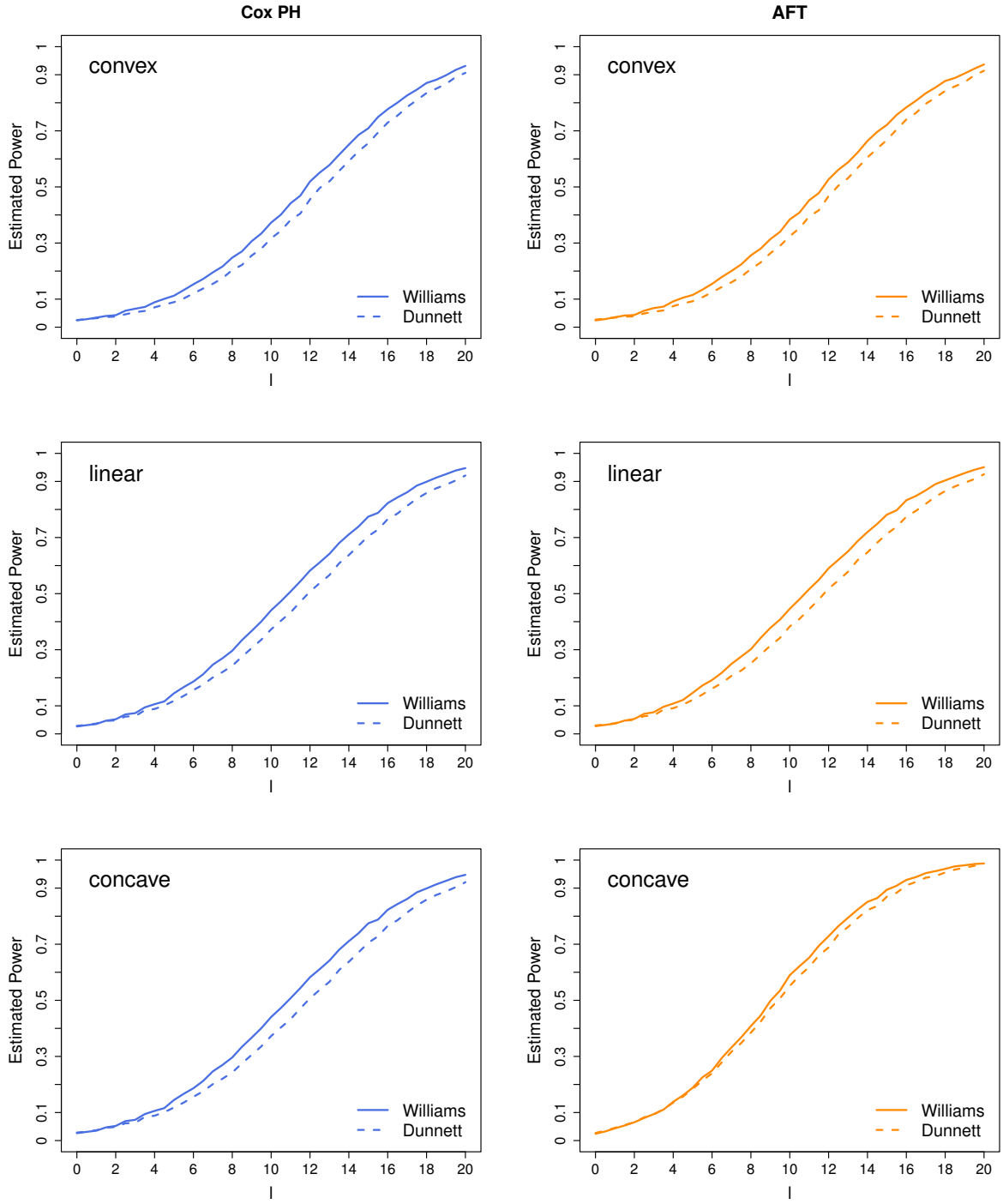


Figure 4.12: Estimated power of the one-sided 97.5% simultaneous confidence intervals for Williams-type contrasts versus Dunnett-type contrasts for a convex dose-response shape (upper row), linear dose-response shape (middle row), and concave dose-response shape (bottom row) for the Cox PH model (left column) and the AFT model (right column) in balanced settings with $n_4 = 50$ objects in dosage group 4.

The power of the simultaneous confidence intervals for Williams-type contrasts is considerably higher compared to simultaneous confidence intervals for Dunnett-type contrasts in both survival models and for all three dose-response shapes. The simultaneous confidence intervals for Williams-type contrasts in the AFT model are slightly superior to the corresponding intervals in the Cox PH model especially for concave profiles of the effects. Only in the AFT model the power to detect a trend is higher when the dose-response shape is concave compared to the linear or convex shape.

## 4.5 Comparisons of Increasing Dosages of Piperonyl Butoxide with Control

In the NTP-120 study, survival was observed in three groups of female mice: a control group (C) and two dose groups (D1 and D2) treated with increasing doses of piperonyl butoxide. There were 50 observations in each dose group and 20 observations in the control. Many observations were censored, mostly due to mice being killed at day 784 (dose groups) and 826 (control) (National Toxicology Program (U.S.), 1979). The frequencies of events, early censoring, and scheduled sacrifices for the three groups are given in Table 4.7.

| | C | D1 | D2 |
|---|---|---|---|
| Events | 5 | 12 | 16 |
| Early Censored | 0 | 3 | 0 |
| Scheduled Sacrifice | 15 | 35 | 34 |

Table 4.7: Censored and uncensored deaths in the NTP-120 study on piperonyl butoxide in female mice.

The approach recommended by the NTP for analysis of mortality in long-term carcinogenicity bioassays is Tarone's life table test for a linear dose-related trend (Tarone, 1975). This test gives a $p$-value of 0.053, i.e., a linear trend cannot be demonstrated. However, the $p$-value is close to the $\alpha = 0.05$ level and the question arises whether the a-priori assumption of a linear trend is plausible or whether a biologically relevant monotonic reduction of survival occurs when Williams-type hazard ratios are used as effect size. The simultaneous confidence limits from the Williams-type procedure take the uncertainty of estimation into account, as well as the multiplicity adjustment, which is required due to considering more than one dose-response shape. The hazard ratio and its lower confidence limit allow a more appropriate interpretation than a $p$-value: both statistical significance and biological relevance can be assessed. Alternatively, the survival time ratio with associated upper confidence limit can be used. In the following, the data are analyzed using both the approach from the Cox PH model and from the AFT model.

The Williams-type hazard ratios were 3.83 when comparing dose group D2 with the control group, and 3.18 when comparing the merged sample of all treated animals (D1, D2) with the untreated controls. Both associated 97.5% lower simultaneous confidence

intervals included the value 1 (Table 4.8). Thus, no significantly increased hazard of mortality was present with increasing dosage. The largest lower confidence limit of 0.82 is so small, that a biological relevant increase which is statistically non-significant is rather unlikely.

| Comparison | Estimated Hazard Ratio | 97.5% Interval |
|---|---|---|
| C vs. D2 | 3.83 | $[0.82, \infty)$ |
| C vs. (D1, D2) | 3.18 | $[0.71, \infty)$ |

Table 4.8: 97.5% lower simultaneous confidence intervals for Williams-type hazard ratios from a Cox PH model for comparison of piperonyl butoxide dosages with control in female mice.

When time of survival was modeled by an AFT model the Williams-type survival time ratios were estimated to be 0.81 and 0.87 (Table 4.9). Both associated simultaneous upper confidence intervals included a survival time ratio of 1, leading to the same conclusion that no significant trend in mortality was present with increasing dosage.

| Comparison | Estimated Survival Time Ratio | 97.5% Interval |
|---|---|---|
| C vs. D2 | 0.81 | $(0, 1.24]$ |
| C vs. (D1, D2) | 0.86 | $(0, 1.30]$ |

Table 4.9: 97.5% upper simultaneous confidence intervals for Williams-type survival time ratios from an AFT model for comparison of piperonyl butoxide dosages with control in female mice.

## 4.6 Summary

Multiple comparisons with a control are a common issue in biological and medical studies with survival endpoint. Simultaneous inference in survival models can be performed according to the procedure by Hothorn et al. (2008) based on the asymptotic normality of contrasts of the parameter estimates, i.e. (penalized) partial likelihood estimates in Cox models and maximum-likelihood estimates in AFT models.
One-sided simultaneous confidence intervals for Dunnett-type contrasts in the frailty Cox model can serve as a tool for analyzing multicenter clinical trials which compare several experimental therapies with a standard therapy under adjustments for covariate information, if applicable, and taking the multicenter structure into account. Simulations showed that the joint coverage probability for many-to-one comparisons of treatment effects is close to the nominal confidence level $1 - \alpha$ even for relatively small sample sizes, both in balanced and in various unbalanced designs. Especially in the settings with a large number of clusters with few observations within each cluster, which often occurs in multicenter clinical trials, the joint coverage probability of the simultaneous confidence intervals equals the nominal confidence level. The systematical decrease of the coverage

probability with increasing distance of the effect size from the null hypothesis results from bias in the penalized partial likelihood estimates. However, for the effect sizes commonly achieved in clinical trials the undercoverage is not relevant.
One-sided simultaneous confidence intervals for Williams-type contrasts in Cox PH models or AFT models can be employed for trend analyses in toxicological studies comparing increasing dosages of a substance with a negative control. Williams-type contrasts take the order restriction of the alternative into account and are sensitive to several dose-response shapes. Simulations under the null hypothesis showed a joint coverage close to the nominal confidence level in the balanced setting, which is recommended by the National Toxicology Program, already for sample sizes of 30 in each dose group. However, in designs with fewer observations in the control group, which was used in some previous studies, the procedure is slightly conservative and less powerful than in the balanced case. Simultaneous inference using Williams-type contrasts is more powerful than using Dunnett-type contrasts for investigation of dose-related trends in mortality, agreeing with the results of simulations for normal endpoints by Bretz and Hothorn (2000).
Since the procedures evaluated in this chapter show good performance regarding the coverage of simultaneous confidence intervals under both the null hypothesis and the alternative their use in the described applications from biological and medical research. A further benefit of simultaneous confidence intervals in survival models is the easy transformation to simultaneous confidence intervals for effects sizes (hazard ratios or survival time ratios) which can be interpreted regarding clinical importance or biological relevance.

# Chapter 5

# Multiple Comparisons in Semiparametric Mixed Models

In this chapter an approach for multiple curve comparisons and an extension for multiple comparisons of areas under the curve are proposed. Curves are modeled by a semiparametric mixed model estimated in the linear mixed model framework. For the multiple comparisons the simultaneous inference procedure by Hothorn et al. (2008) is employed to test hypotheses on the equality of curves or areas thereunder specified by multiple contrasts in the linear mixed model representation of the model. The aim of this chapter is to evaluate the small-sample performance of these methods.

In many biological experiments data arise as curves, e.g. growth curves (Krafty, Gimotty, Holtz, Coukos, and Guo, 2008), hormone level profiles (Zhang et al., 2000), drug concentration profiles (Bertrand, Comets, Chenel, and Metré, 2012), and antigen trajectories (Bhadra, Daniels, Kim, Ghosh, and Mukherjee, 2012). The development of the proposed method is motivated by a study that examines the impact of mutations of the EphA4 gene on the dorsal funiculus, a morphological structure in the spinal cord. If the EphA4 gene is completely conserved, the length of the dorsal funiculus forms a characteristic, nonlinear curve over the spinal cord. Mutations of EphA4 in which different domains of the gene are knocked out lead to reduction of the dorsal funiculus and a modified form of the curve. Figure 5.1 displays measurements of the dorsal funiculus at 25 positions over the spinal cord for three groups of mice characterized by different variants of the EphA4 gene. The data will be discussed in greater detail in Subsection 5.3. To study the biological function of the knocked out domains of EphA4 the dorsal funiculus curves are to be compared between three genotypes, which correspond to two mutated genotypes and a wildtype genotype. In addition to detecting an overall difference between two curves, it is of interest to investigate which region of the spinal cord is sensitive to lack of the EphA4 domains, i.e., in which positions the curves differ between the EphA4 genotypes.

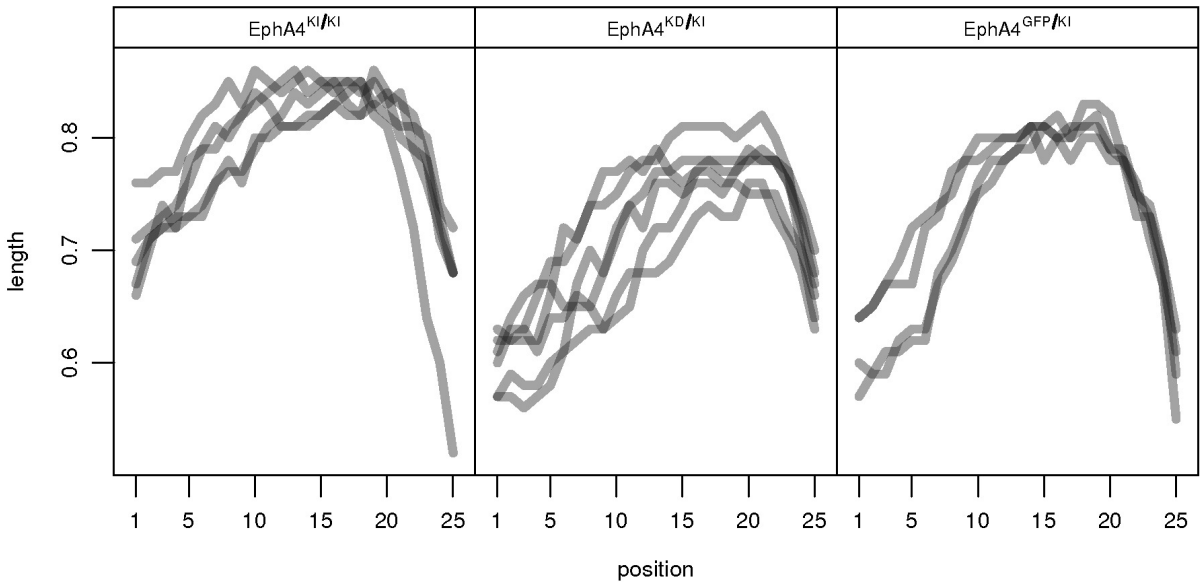


Figure 5.1: Length of the dorsal funiculus at 25 positions for each mouse for EphA4 genotypes KI/KI (left), GFP/KI (center), and KD/KI (right).

This testing problem, which addresses comparisons of several group-specific curves by comparing pairs of curves over a grid along the curves for several pairs of curves, will in the following be referred to as multiple curve comparisons. These comparisons can be Dunnett-type comparisons, where the curve of a control group is compared to the curves of several other groups, Tukey-type comparisons, where all possible pairs of groups are compared, or any other kind of multiple comparisons.
Pairwise comparisons of several curves over a grid along the curves result in a multiple testing problem with the total number of tests equal to the number of pairwise comparisons of two groups multiplied with the number of positions, in which the curves are to be compared. Neglecting the multiplicity and testing each comparison on a level of $\alpha$ leads to a probability of false positive findings increased above the nominal level.
In similar setups, Zhang et al. (2000) used nonparametric functions to model smooth time effects on hormone data and proposed a scaled $\chi^2$-test statistic based on the fitted group-level curves to examine an overall difference between the curves of two groups. The procedure was extended by Kong and Yan (2011) to the overall comparison of more than two groups. After detection of an overall difference between any curves pairwise group comparisons with multiplicity adjustment using the Bonferroni method are suggested. Behseta and Chenouri (2011) modeled smooth intensity functions of groups of neurons by Bayesian adaptive regression splines. In order to compare the curves belonging to two different experimental conditions they developed a parametric approach using a modified Hotelling $T^2$ statistic and a nonparametric approach based on a signed-rank test statistic. However, all existing approaches do not provide information on the positions in which the curves differ if an overall difference exists.

In this chapter an approach for multiple curve comparisons is derived by combination of two frameworks. The first one exploits the connection between semiparametric mixed models and linear mixed models (Fahrmeir, Kneib, and Lang, 2004). Smooth curves of unknown form for several groups are nonparametrically modeled using penalized splines to describe a smooth curve for each group. Random effects are used to model the subject-specific deviation from the group-level curve leading to a semiparametric mixed model. Asymptotic normal parameter estimates can be obtained by first representing the semiparametric mixed model by a linear mixed model and then using best linear unbiased prediction (BLUP).
The framework for simultaneous inference in general parametric models by Hothorn et al. (2008) can then be used for multiple comparisons of the estimated curves or the areas thereunder. For multiple curve comparisons, multiple contrasts of parameters from the linear mixed model are built such that each contrast represents the difference of two curves at a particular position over the curve with the set of contrasts defining all necessary single comparisons. Adjusted $p$-values for each single comparison can be calculated based on the asymptotic normality of the estimated contrasts following the approach by Hothorn et al. (2008).

Pharmacological studies often measure the concentrations of a compound in blood plasma repeatedly after drug exposition and compare the concentrations between different doses or administration forms over time. However, not the concentration-time curves themselves are compared, but the areas thereunder which are used as a measure for the total body exposure to drug after administration of one dose. One approach to estimate the area under a curve is to apply the trapezoidal rule on the means of measurements at each time point. Bailer (1988) describes a method for pairwise comparisons of AUCs estimated by these linear combinations of the sample means under the independence assumption for the measurements and suggests the application of the Bonferroni method if more than two AUCs are to be compared. In study designs where measurements at each time point are available for all subjects, the AUC is usually estimated by the trapezoidal rule for each subject and multiple comparisons of group-level AUCs are performed using ANOVA (Westlake, 1973).
The second part of this chapter extends the approach for multiple curve comparisons to multiple comparisons of AUCs. Again, the curves are described by a semiparametric mixed model and estimated within the linear mixed model representation. Multiple contrasts in this model are then defined such that each contrast represents the difference of the AUCs of two groups with the set of contrasts defining the multiple AUC comparisons of interest. Differences between the AUCs can be inspected by adjusted $p$-values for each comparison of two AUCs according to the approach by Hothorn et al. (2008).

Both approaches are evaluated using simulation studies. The small sample behavior receives special attention as most studies with the objective of comparing several curves or the areas thereunder study only few subjects in each group and a limited number of measurements for each subject. We explore the performance for Dunnett- and Tukey-type multiple curve comparisons. The familywise error rate is estimated in small samples when curves not differing are pairwise compared over a grid along the curves. The power

to detect differences between curves is investigated when one curve partly differs from the others. Additionally, the performance of the procedure is explored when parts of the curves are missing due to censoring.
The relevance of the method for multiple curve comparisons is then demonstrated by comparing curves describing the formation of the dorsal funiculus between two mutated and a wildtype genotype of the EphA4 gene in mice to detect which regions of the spinal cord are affected by lack of certain domains of this gene.
The approach for multiple comparisons of areas under the curve is complicated by an additional estimation, which is the approximation of the area employing the trapezoidal rule. This might cause impairment of the multiple AUC comparison properties. The extent of the problem is gauged by simulations for multiple comparisons of AUCs in small samples of various designs using both the familywise error rate and power as criteria.
A reanalysis of a study comparing the exposure to benzene between three dosages of the compound illustrates the method. The comparison is performed by comparing the areas under the associated concentration-time curves.

## 5.1 Multiple Comparisons of Curves and Areas Under the Curves

### 5.1.1 Model

Let $K$ be the number of groups with $N(k)$ subjects in group $k$, $k = 1, \ldots, K$. For the $i$th subject in the $k$th group, we have measurements $y_{jik}$ taken at positions or time points $x_{jik}$, $j = 1, \ldots, J(ik)$, leading to $N = \sum_{k=1}^{K} \sum_{i=1}^{N(k)} J(ik)$ observations of $y_{jik}$ in total. We assume that for each group $k$, the dependent variable follows a smooth, unknown function $f_k(x)$. We specify a semiparametric mixed model

$$y_{jik} = f_k(x_{jik}) + \alpha_{ik} + \varepsilon_{jik}, \tag{5.1}$$

where the curve of the $i$th subject in the $k$th group is shifted from the groupwise effect function $f_k$ by a random, subject-specific value $\alpha_{ik}$. The homoscedastic errors $\varepsilon_{jik} \sim N(0, \sigma_\varepsilon^2)$ are normal at each point $x_{jik}$.

We approximate the smooth functions $f_k(x)$ by a spline, i.e., a linear combination of $L$ basis functions $B_l : \mathbb{R} \longrightarrow \mathbb{R}^+$ with coefficients $\beta_{kl}$:

$$f_k(x) \approx \sum_{l=1}^{L} B_l(x)\beta_{kl}$$

The model then becomes

$$y_{jik} = \sum_{l=1}^{L} B_l(x_{jik})\beta_{kl} + \alpha_{ik} + \varepsilon_{jik},$$

or in matrix notation

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

The response vector

$$\boldsymbol{y} = (y_{jik}) \in \mathbb{R}^{N\times 1}$$

contains the dependent measurements of all subjects at all time points or positions,

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}^* & & \\ & \ddots & \\ & & \boldsymbol{B}^* \end{pmatrix} \in \mathbb{R}^{N\times(KL)}$$

is a block-diagonal B-spline design matrix with block matrices

$$\boldsymbol{B}^* = \begin{pmatrix} B_1(x_{11k}) & \cdots & B_L(x_{11k}) \\ \vdots & \vdots & \vdots \\ B_1(x_{J(N(k),k),N(k),k}) & \cdots & B_L(x_{J(N(k),k),N(k),k}) \end{pmatrix} \in \mathbb{R}^{\left(\sum_{i=1}^{N(k)} J(ik)\right)\times L},$$

the vector

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{KL\times 1} \quad \text{with} \quad \boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kL}) \in \mathbb{R}^{L\times 1}$$

contains the spline effects such that

$$\sum_{l=1}^{L} B_l(x)\beta_{kl} \approx f_k(x),$$

the vector

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{11}, \ldots, \boldsymbol{\alpha}_{N(k)K}) \in \mathbb{R}^{N\times 1} \quad \text{with} \quad \boldsymbol{\alpha}_{ik} = (\alpha_{ik}) \in \mathbb{R}^{J(ik)\times 1}$$

contains the random, subject-specific deviations from the group-level curve for all subjects, and the vector

$$\boldsymbol{\varepsilon} = (\varepsilon_{jik}) \in \mathbb{R}^{N\times 1}$$

contains the normal, homoscedastic errors of all measurements.

Smoothness of the curves is ensured by introducing a penalty on the spline coefficients $\boldsymbol{\beta}$, leading to the penalized least squares criterion

$$\underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmin}} \left( ||\boldsymbol{y} - (\boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\alpha})||^2 + \sum_{k=1}^{K} \lambda_k \boldsymbol{\beta}^\top \boldsymbol{P}_k \boldsymbol{\beta} + \lambda_{K+1} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right), \tag{5.2}$$

where

$$\boldsymbol{P}_k = \begin{pmatrix} \boldsymbol{0} & & & \\ & \ddots & & \\ & & \boldsymbol{P} & \\ & & & \ddots \\ & & & & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{KL\times KL}$$

are block-diagonal penalty matrices with the $k$th block equal to $\boldsymbol{P} = \boldsymbol{K}^\top \boldsymbol{K} \quad \forall k = 1, \ldots, K$, and $\boldsymbol{K}$ is the second-order differences matrix (Eilers and Marx, 1996).

We now reparameterize the semiparametric mixed model by decomposing the spline coefficients $\boldsymbol{\beta}$ into an unpenalized and a penalized part with unpenalized coefficients $\boldsymbol{\gamma}$ and penalized coefficients $\boldsymbol{\delta}$ as described in Fahrmeir et al. (2004):

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{B}(\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{V}\boldsymbol{\delta}) + \boldsymbol{I}_N\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{B}\boldsymbol{U}\boldsymbol{\gamma} + (\boldsymbol{B}\boldsymbol{V} \mid \boldsymbol{I}_N)(\boldsymbol{\delta}, \boldsymbol{\alpha}) + \boldsymbol{\varepsilon} \\ &= \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \end{aligned} \tag{5.3}$$

with identity matrix $\boldsymbol{I}_N \in \mathbb{R}^{N \times N}$.

The penalized least squares criterion (5.2) then becomes

$$\underset{\boldsymbol{\gamma}, \boldsymbol{\xi}}{\operatorname{argmin}} \left( ||\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{Z}\boldsymbol{\xi})||^2 + \sum_{k=1}^{K+1} \lambda_k \, \boldsymbol{\xi}_k^\top \boldsymbol{\xi}_k \right),$$

where

$$\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{K+1})$$

are the transformed parameters with entries $\boldsymbol{\xi}_k = \boldsymbol{\delta}_k$, $k = 1, \ldots, K$, and $\boldsymbol{\xi}_{K+1} = \boldsymbol{\alpha}$.

According to Ruppert, Wand, and Carroll (2003) the solution of this minimization problem is equivalent to the best linear unbiased prediction (BLUP) estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ in the linear mixed model (5.3) with fixed effects $\boldsymbol{\gamma}$, random effects $\boldsymbol{\xi} \sim N(\boldsymbol{0}, \Xi)$ with $\Xi$ a block diagonal covariance matrix with fixed variances $\sigma^2_{\xi_k} = \sigma^2_\varepsilon / \lambda_k$, and errors $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2_\varepsilon \boldsymbol{I}_N)$ for given $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{K+1})$. Estimates of $\boldsymbol{\beta}$ can then be obtained via BLUP estimation in the linear mixed model (5.3) and the smoothing parameters $\lambda_k$ can be chosen as estimates of the variance ratios $\sigma^2_\varepsilon / \sigma^2_{\xi_k}$ obtained via ML- or REML-methodology.

For multiple tests of hypotheses on linear combinations of the parameters of a linear mixed model, the simultaneous inference procedure of Hothorn et al. (2008) can be used. The application of the method for multiple comparisons of curves fitted by model (5.1) is described in the following section.

### 5.1.2 Hypotheses for Multiple Curve Comparisons

We are looking at $M$ pairwise comparisons of curves, where two groups $k$ and $k'$ are compared in the $m$th hypothesis

$$\begin{aligned} \mathrm{H}^0_m : \sup_{x \in \mathbb{R}} |f_k(x) - f_{k'}(x)| = 0, \\ 1 \leq k < k' \leq K, \quad m = 1, \ldots, M. \end{aligned}$$

We approximate these hypotheses by comparing the associated splines on a grid $\{x^1, \ldots, x^S\}$

$$\begin{aligned} \mathrm{H}^0_{m,x} : (B_1(x), \ldots, B_L(x))(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}) = 0 \\ \forall x \in \{x^1, \ldots, x^S\}, \quad m = 1, \ldots, M, \end{aligned}$$

with the grid values being the positions of the measurements. These hypotheses can be reformulated to

$$\mathrm{H}^0_{m,x} : \boldsymbol{C}_{m,x}\boldsymbol{\beta} = 0 \quad \forall x \in \{x^1, \ldots, x^S\}, \quad m = 1, \ldots, M,$$

using

$$\begin{aligned} (B_1(x), \ldots, B_L(x))(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}) &= \\ \underbrace{(B_1(x), \ldots, B_L(x))\boldsymbol{D}_m}_{=: \boldsymbol{C}_{m,x} \in \mathbb{R}^{1 \times KL}}\boldsymbol{\beta} &= \boldsymbol{C}_{m,x}\boldsymbol{\beta}, \end{aligned}$$

with $\boldsymbol{D}_m = (\boldsymbol{0} \,|\, \overbrace{\boldsymbol{I}_L}^{k\text{th block}} \,|\, \boldsymbol{0} \,|\, \overbrace{-\boldsymbol{I}_L}^{k'\text{th block}} \,|\, \boldsymbol{0}) \in \mathbb{R}^{L \times KL}$.

The hypotheses for the $M$ pairwise comparisons of curves over all positions $x$ can then be specified by

$$\mathrm{H}^0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0},$$

with $\boldsymbol{C} \in \mathbb{R}^{MS \times KL}$ denoting the row stack of $\boldsymbol{C}_{m,x}$, $x = x^1, \ldots, x^S$, $m = 1, \ldots, M$.

### 5.1.3 Hypotheses for Multiple AUC Comparisons

We are looking at $M$ pairwise comparisons of areas under curves, where two groups $k$ and $k'$ are compared in the $m$th hypothesis

$$\begin{aligned} \mathrm{H}^0_m : \int f_k(x)\,dx - \int f_{k'}(x)\,dx = 0, \\ 1 \leq k < k' \leq K, \quad m = 1, \ldots, M. \end{aligned}$$

We estimate the integrals by the trapezoidal rule

$$\int_{x^1}^{x^S} f(x) \approx \frac{1}{2} \cdot \sum_{s=1}^{S-1} (x^{s+1} - x^s) \cdot (f(x^{s+1}) + f(x^s))$$

on a grid $\{x^1, \ldots, x^S\}$ leading to the approximated hypotheses

$$\begin{aligned} \mathrm{H}^0_m : \frac{1}{2} \cdot \sum_{s=1}^{S-1} (x^{s+1} - x^s) \cdot (f_k(x^{s+1}) + f_k(x^s)) - \frac{1}{2} \cdot \sum_{s=1}^{S-1} (x^{s+1} - x^s) \cdot (f_k(x^{s+1}) + f_k(x^s)) = 0, \\ 1 \leq k < k' \leq K, \quad m = 1, \ldots, M, \end{aligned}$$

and model the unknown functions $f_k$ and $f_{k'}$ by the associated splines leading to

$$\mathrm{H}_m^0 : \frac{1}{2} \cdot \left( \begin{array}{rcl} & (x^2 - x^1) & \cdot \quad (B_1(x^1), \ldots, B_L(x^1)) \\ + & \sum_{s=2}^{S-1} (x^{s+1} - x^{s-1}) & \cdot \quad (B_1(x^s), \ldots, B_L(x^s)) \\ + & (x^S - x^{S-1}) & \cdot \quad (B_1(x^S), \ldots, B_L(x^S)) \end{array} \right) (\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}) = 0$$

$$1 \leq k < k' \leq K, \quad m = 1, \ldots, M.$$

These hypotheses can be reformulated to

$$\mathrm{H}_{m,x}^0 : \boldsymbol{C}_m \boldsymbol{\beta} = 0, \quad , \quad m = 1, \ldots, M,$$

using

$$\frac{1}{2} \cdot \left( \begin{array}{rcl} & (x^2 - x^1) & \cdot \quad (B_1(x^1), \ldots, B_L(x^1)) \\ + & \sum_{s=2}^{S-1} (x^{s+1} - x^{s-1}) & \cdot \quad (B_1(x^s), \ldots, B_L(x^s)) \\ + & (x^S - x^{S-1}) & \cdot \quad (B_1(x^S), \ldots, B_L(x^S)) \end{array} \right) (\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}) \;=\;$$

$$\underbrace{\frac{1}{2} \cdot \left( \begin{array}{rcl} & (x^2 - x^1) & \cdot \quad (B_1(x^1), \ldots, B_L(x^1)) \\ + & \sum_{s=2}^{S-1} (x^{s+1} - x^{s-1}) & \cdot \quad (B_1(x^s), \ldots, B_L(x^s)) \\ + & (x^S - x^{S-1}) & \cdot \quad (B_1(x^S), \ldots, B_L(x^S)) \end{array} \right)}_{=: \boldsymbol{C}_m \in \mathbb{R}^{1 \times KL}} \boldsymbol{D}_m \boldsymbol{\beta} \;=\; \boldsymbol{C}_m \boldsymbol{\beta}$$

with $\boldsymbol{D}_m = (\boldsymbol{0} \,|\, \overbrace{\boldsymbol{I}_L}^{k\text{th block}} \,|\, \boldsymbol{0} \,|\, \overbrace{-\boldsymbol{I}_L}^{k'\text{th block}} \,|\, \boldsymbol{0}) \in \mathbb{R}^{L \times KL}$.

The hypotheses for the $M$ pairwise comparisons of areas under the curves can then be specified by

$$\mathrm{H}^0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0},$$

with $\boldsymbol{C} \in \mathbb{R}^{M \times KL}$ the row stack of $\boldsymbol{C}_m$, $m = 1, \ldots, M$.

If the grid values are chosen as the positions of the measurements $\boldsymbol{C}_m\boldsymbol{\beta}$ approximates

$$\int f_k(x)\, d\mathbb{P}_X - \int f_{k'}(x)\, d\mathbb{P}_X,$$

whereas if the grid values are chosen on a dense grid the $\boldsymbol{C}_m\boldsymbol{\beta}$ approximates

$$\int f_k(x)\, dx - \int f_{k'}(x)\, dx.$$

### 5.1.4 Inference

The BLUP estimates $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$ asymptotically follow a multivariate normal distribution

$$\sqrt{n}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}) - \mathbb{E}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}))) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma} = \mathbb{V}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}))$ (Ruppert et al., 2003).

The covariance of $\hat{\boldsymbol{\beta}} = \boldsymbol{U}\hat{\boldsymbol{\gamma}} + \boldsymbol{V}\hat{\boldsymbol{\delta}}$ can be calculated as

$$\begin{aligned}
\mathbb{V}(\hat{\boldsymbol{\beta}}) &= \mathbb{V}(\boldsymbol{U}\hat{\boldsymbol{\gamma}} + \boldsymbol{V}\hat{\boldsymbol{\delta}}) \\
&= \mathbb{V}(\underbrace{[(\boldsymbol{U} \mid \boldsymbol{0}) + (\boldsymbol{0} \mid \boldsymbol{V})]}_{=:\boldsymbol{W}}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})) \\
&= \boldsymbol{W}\mathbb{V}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}))\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top \quad \text{(Fahrmeir et al., 2004).}
\end{aligned}$$

Therefore we get

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \underbrace{\mathbb{E}(\hat{\boldsymbol{\beta}})}_{\neq \boldsymbol{\beta}}) \xrightarrow{d} N(0, \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top),$$

$$\sqrt{n}(\boldsymbol{D}_m\hat{\boldsymbol{\beta}} - \underbrace{\boldsymbol{D}_m\mathbb{E}(\hat{\boldsymbol{\beta}})}_{=0}) \xrightarrow{d} N(0, \boldsymbol{D}_m\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top\boldsymbol{D}_m^\top),$$

and

$$\sqrt{n}\,\boldsymbol{C}\hat{\boldsymbol{\beta}} \xrightarrow{d} N(0, \boldsymbol{C}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top\boldsymbol{C}^\top).$$

The covariance matrix $\boldsymbol{\Sigma}$ of the estimates $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$ can be estimated by a Bayesian posterior covariance matrix $\hat{\mathbb{V}}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}))$ (Silverman, 1985).

With

$$\sqrt{n}\,\hat{\mathbb{V}}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})) \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}$$

we get

$$\sqrt{n}\,\boldsymbol{C}\boldsymbol{W}\hat{\mathbb{V}}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}))\boldsymbol{W}^\top\boldsymbol{C}^\top \xrightarrow{\mathbb{P}} \sqrt{n}\,\boldsymbol{C}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top\boldsymbol{C}^\top.$$

With the contrast matrix $\boldsymbol{C}$ chosen as described in Subsection 5.1.2 or Subsection 5.1.3, respectively, adjusted $p$-values for multiple comparisons of two curves at all positions, or for multiple comparisons of the difference of the areas under two curves, respectively, can then be computed based on the distribution

$$\sqrt{n}\,\boldsymbol{C}\hat{\boldsymbol{\beta}} \xrightarrow{d} N(0, \boldsymbol{C}\boldsymbol{W}\hat{\mathbb{V}}((\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}))\boldsymbol{W}^\top\boldsymbol{C}^\top)$$

as described in Section 2.4.

## 5.2 Behavior of Multiple Curve Comparisons

The following simulations estimate the FWER and power of multiple curve comparisons using the proposed approach in small samples and for different combinations of number of subjects per group and number of measurements per subject. A further objective is to compare the influence of the measurement spacing on the properties of multiple curve comparisons. Additionally, the impact of censoring is studied.

### 5.2.1 Simulation Setup

The FWER and the power of the testing procedure are estimated for Dunnett- and Tukey-type comparisons of three curves.
For $N(k)$ subjects in each group $k$, observations at $J(ik)$ values $x \in [0, 1]$ are simulated from the "true" function

$$f(x) = x^{11} \cdot (10 \cdot (1 - x))^6 + 10 \cdot (10x)^3 \cdot (1 - x)^{10} \tag{5.4}$$

scaled to the interval $[0, 1]$, with a subject-specific error $\alpha_{ik} \sim N(0, \sigma_\alpha^2)$ and a random error $\varepsilon_{jik} \sim N(0, \sigma_\varepsilon^2)$ added to each observation:

$$y_{jik} = f(x_{jik}) + \alpha_{ik} + \varepsilon_{jik}. \tag{5.5}$$

The function $f$ was taken from the simulations in Wood (2006b) and is displayed in Figure 5.2 (black curve).

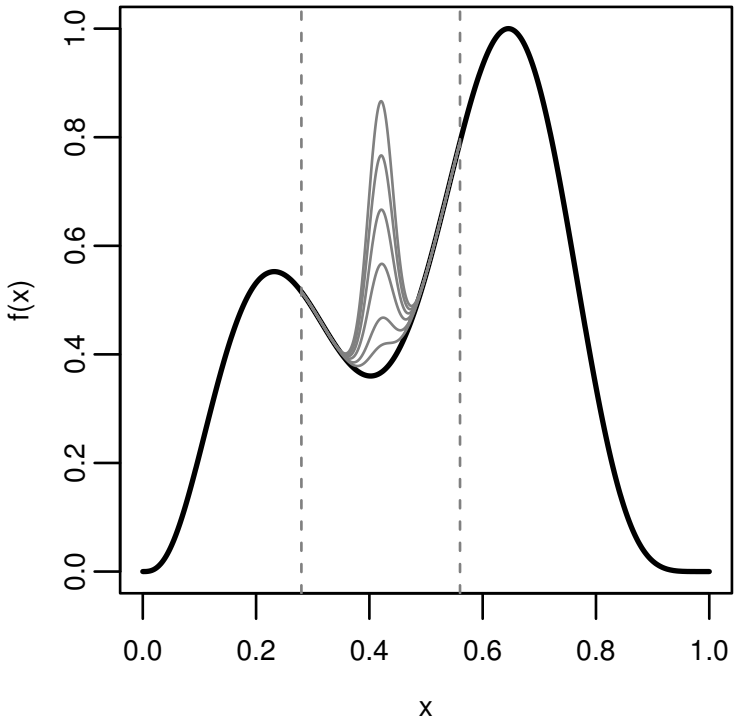


Figure 5.2: True smooth function for the estimation of the familywise error rate (black) and smooth function of group 3 for values of $a$ varying between 0 and 0.5 for the estimation of power (gray).

The standard deviations for the error terms and the random effects were chosen in three different combinations:

- $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.02$,
- $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.05$,
- $\sigma_\alpha = 0.05$ and $\sigma_\varepsilon = 0.02$.

Three different grid patterns for $x$ were considered:

(a) equally spaced on $[0, 1]$,

(b) continuous uniformly distributed on $[0, 1]$ with different positions for different subjects,

(c) decreasing density of $x$ (positions at the quantiles of the Beta(1,3) distribution).

We investigated scenarios with 15, 20, or 25 positions and 5, 10, 15, or 20 subjects per group.

For the estimation of the FWER observations were simulated from the "null model" (5.5) for all three groups. The curves were fitted by the semiparametric mixed model (5.1) with the smooth terms approximated by a linear combination of B-splines basis functions (Eilers and Marx, 1996). The fitted curves were compared at each position for settings (a) and (c), and at $N(k)$ equally spaced position for setting (b). The FWER was estimated by the portion of 1,000 datasets in which at least one difference was found among all comparisons made. The same datasets were used for both Dunnett- and Tukey-type comparisons.
Additionally, settings (a), (b) and (c) were examined when rather small observed values of $y_{jik}$ were censored. In practice, if $\alpha_{ik} + \varepsilon_{jik} < -sd(\alpha_{ik} + \varepsilon_{jik})$ for any measurement at the fourth or higher position, i.e., $x_{jik}$, $j > 3$, this observations and all following observations from the same subject were censored until the proportion of censored observations was approximately 25%. The nominal level $\alpha$ was set to 0.05.

To investigate the power of the procedure, the observations of group 3 were simulated from a function $f_3$, which differed from function (5.4) for $x$ values in the interval $[0.28, 0.56]$:

$$f_3(x) = f(x) + \exp\left(-\frac{(x - 0.42)^2}{0.01}\right) \cdot a \cdot I_{[0.28,0.56]}(x), \tag{5.6}$$

with $a \in [0, 0.5]$ to increase the difference between $f_3$ and $f$ (see Figure 5.2). The variances were chosen equally $\sigma_\alpha = \sigma_\varepsilon = 0.02$. For equally spaced positions on $[0, 1]$, the number of positions with values differing between $f$ and $f_3$ are 4, 5, or 7 for 15, 20, or 25 positions in total.
The power of the procedure was estimated by the portion of 1,000 simulated datasets in which at least one significant difference was correctly found (disjunctive power).

### 5.2.2 Simulation Results

#### Familywise Error Rate

The estimated FWER for all simulation scenarios is shown in Figures 5.3, 5.4, and 5.5.

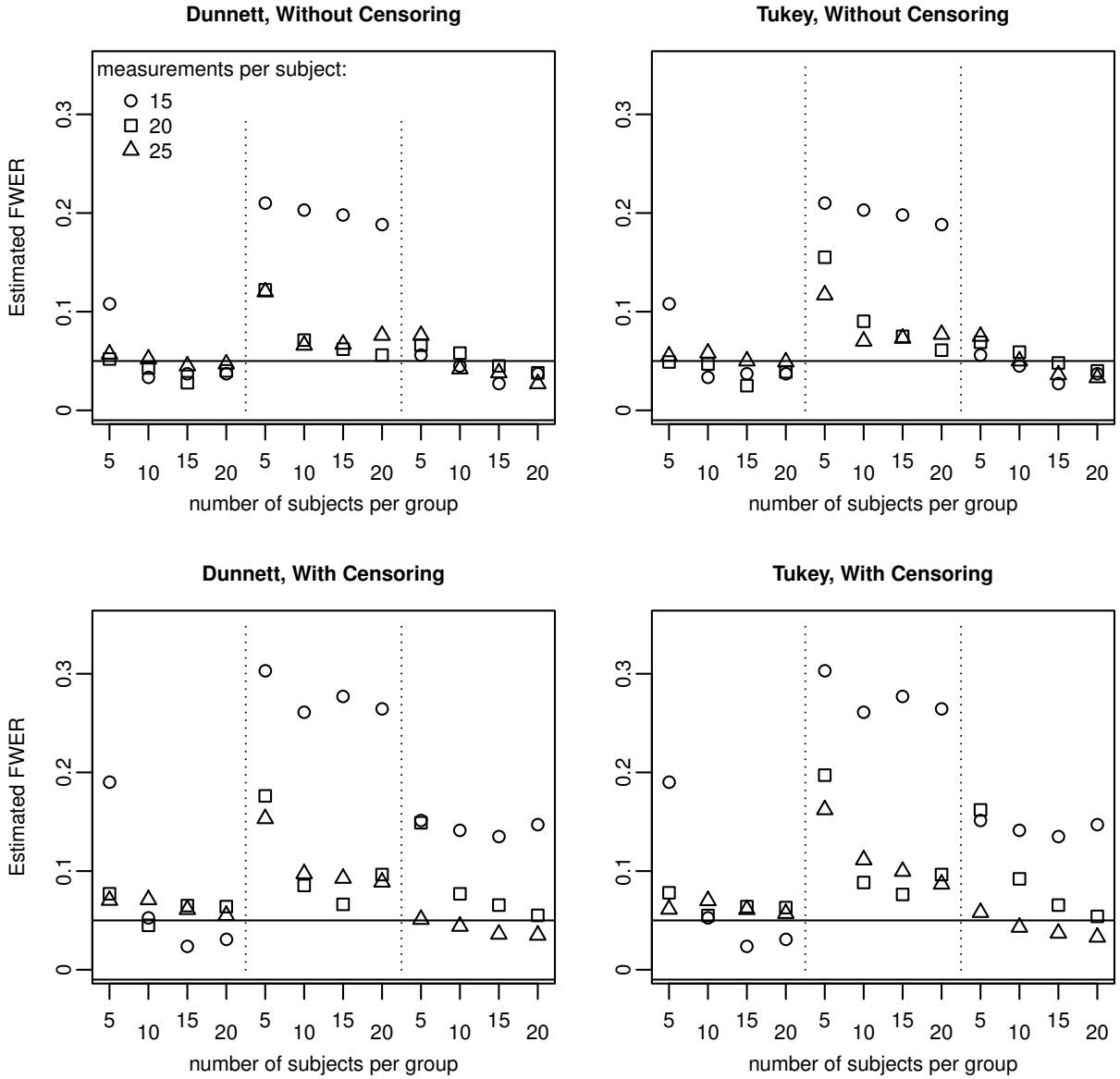


Figure 5.3: Estimated familywise error rate for Dunnett-type (left column) and Tukey-type (right column) comparisons of three curves for setting (a) (left section), (b) (middle section), and (c) (right section) each estimated from 1,000 datasets without censored observations (top row) and with censored observations (bottom row) with standard deviations $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.02$.

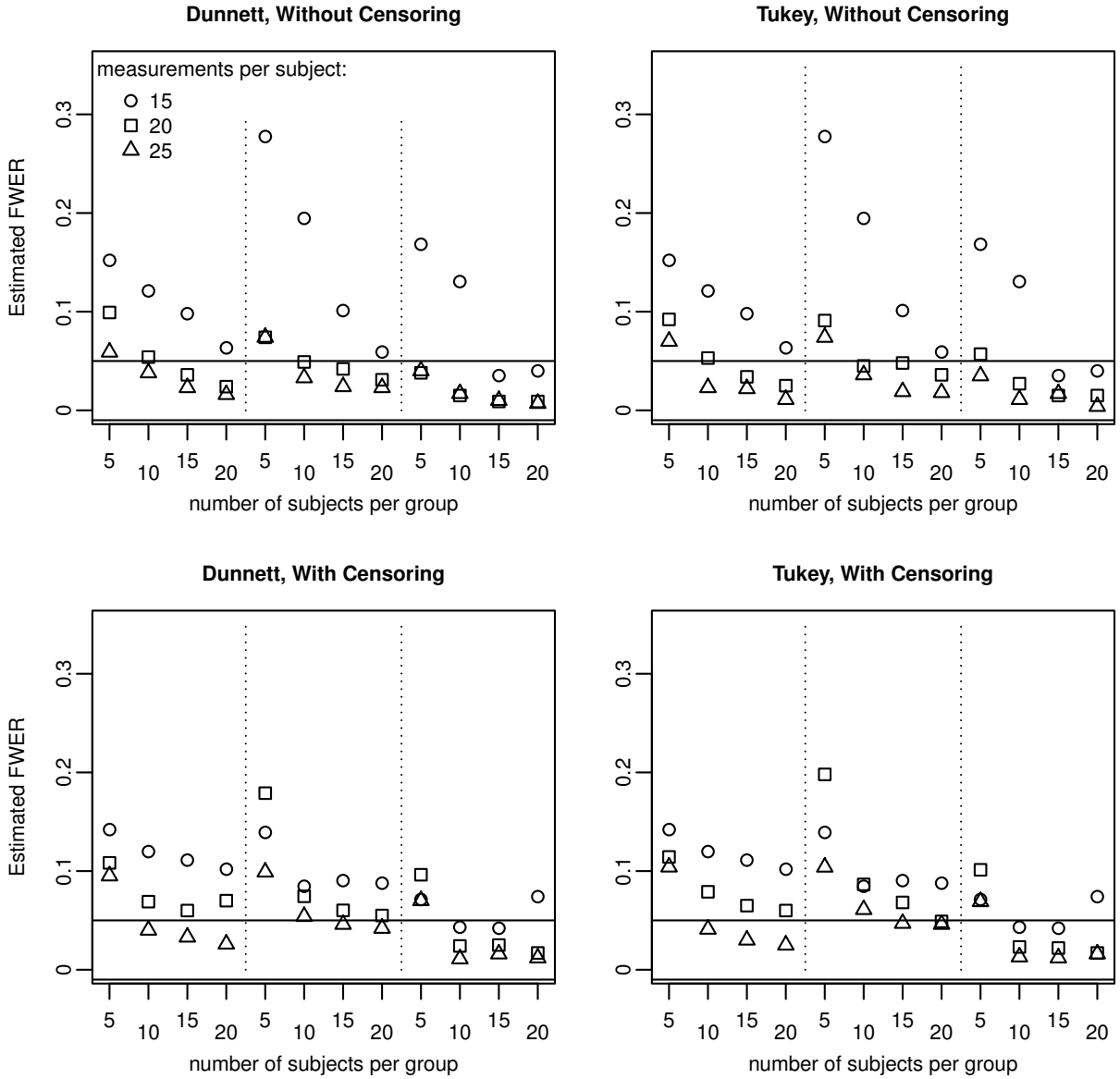


Figure 5.4: Estimated familywise error rate for Dunnett-type (left column) and Tukey-type (right column) comparisons of three curves for setting (a) (left section), (b) (middle section), and (c) (right section) each estimated from 1,000 datasets without censored observations (top row) and with censored observations (bottom row) with standard deviations $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.05$.

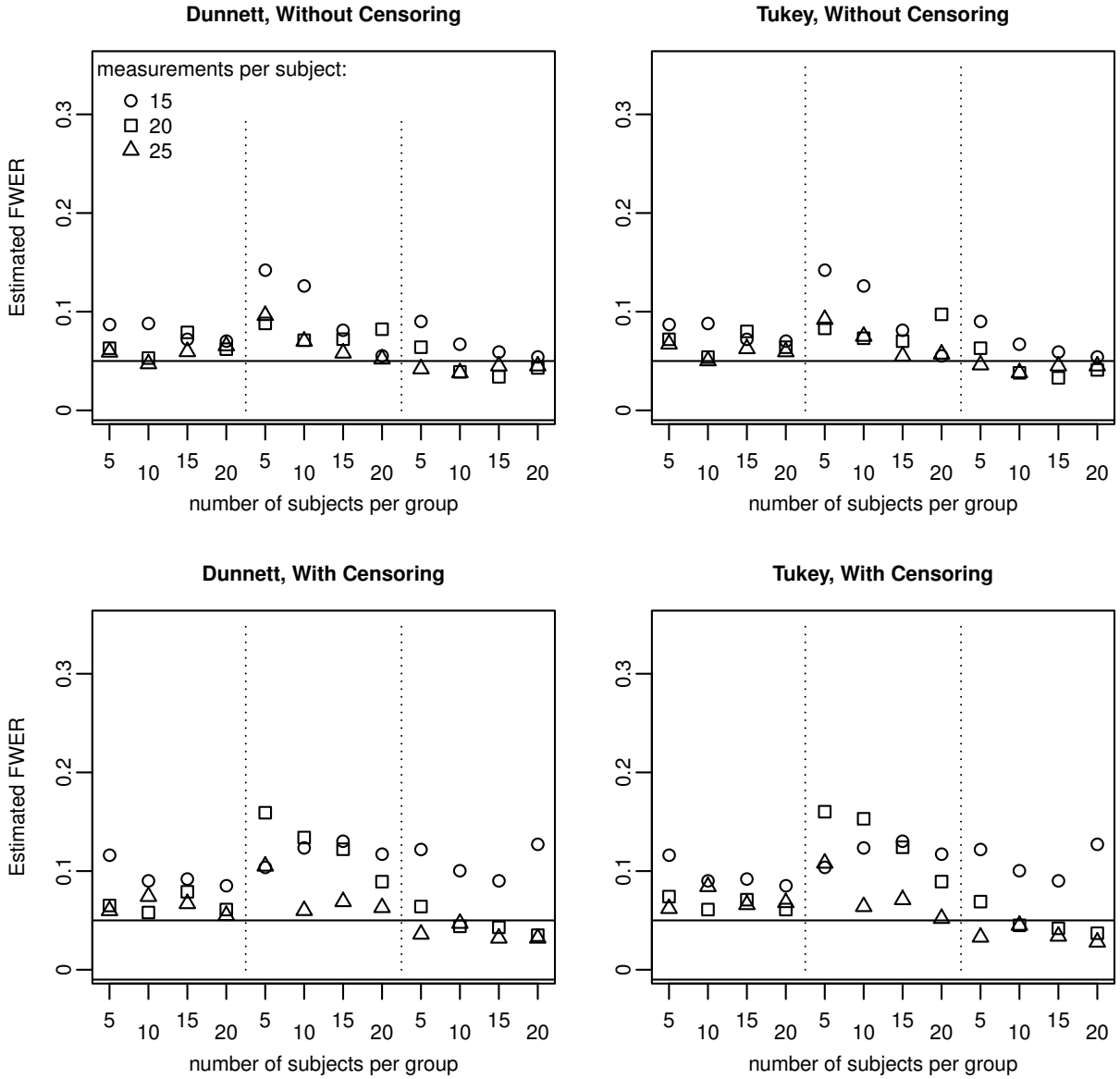


Figure 5.5: Estimated familywise error rate for Dunnett-type (left column) and Tukey-type (right column) comparisons of three curves for setting (a) (left section), (b) (middle section), and (c) (right section) each estimated from 1,000 datasets without censored observations (top row) and with censored observations (bottom row) with standard deviations $\sigma_\alpha = 0.05$ and $\sigma_\varepsilon = 0.02$.

The results are similar for Dunnett- and Tukey-type comparisons, with the estimated FWER being slightly higher for Tukey-type comparisons in most settings. Looking at the designs without censoring and with the same measurement points for each subject (settings (a) and (c)) control of the FWER is achieved when 25 measurements are provided for each subject and the groups contain 5 or more subjects, or when 20 measurements are provided for each subject and the groups contain at least 10 subjects. For comparisons of curves fitted for measurements taken at random positions (setting (b)) the estimated FWER is often above the nominal level. Occurrence of censoring results in an increase of the FWER, but 25 measurements of each subject are sufficient for a FWER close to 0.05. The results differ for the different choices of variances of the error terms and the random effects. The procedure tends to become conservative when the variance for the errors is higher than the variance of the random effects (Figure 5.4), and liberal for the other way round (Figure 5.5). Equal variances $\sigma_\varepsilon^2$ and $\sigma_\alpha^2$ are the combination with the FWER closest to the nominal level.

### Power

The estimated power curves for setting (a), 15 subjects per group, and 15, 20, or 25 positions are shown in Figure 5.6. The power is slightly higher for curves fitted from measurements taken at fewer positions compatible to the estimated FWER for 15 subjects per group, where the procedure gets conservative with increasing number of positions. The power is rather low over a wide range of the parameter $a$ which controls how the curve of the third group differs from the other curves. A considerable difference in the curves is needed for the procedure to detect a difference.

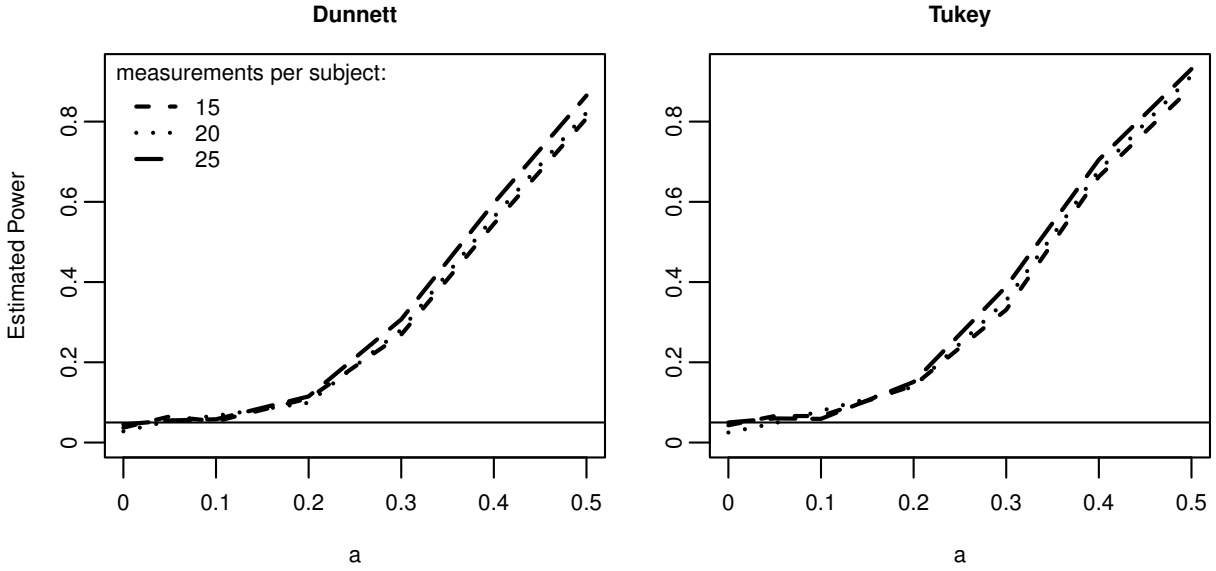


Figure 5.6: Power of Dunnett- and Tukey-type comparisons of three curves where the curve of one group differed from the others according to equation (5.6) for varying values of $a$. The number of subjects per group was 15, and standard deviations were $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.02$.

## 5.3 Comparisons of Dorsal Funiculus Curves Between Variants of the EphA4 Gene

The protein EphA4 plays a major role for the development of the central nervous system. Absence of EphA4 leads to neuronal axons not finding their target cell and neural networks not properly connecting. EphA4 is also required for the development of the so-called dorsal funiculus, a morphological structure in the spinal cord comprised of major axon bundles. When the EphA4 gene is knocked out or enzymatically inactive, formation of the dorsal funiculus is impaired (Kullander, Mather, Diella, Dottori, Boyd, and Klein, 2001; Egea and Klein, 2007).

In wildtype mice with the protein EphA4 completely conserved the length of the dorsal funiculus forms a characteristic nonlinear curve over a subsection of the spinal cord. A parametric model describing the form of the dorsal funiculus size along the spinal cord does not exist. In order to investigate the role of EphA4 mutations on formation of the dorsal funiculus, we compared the dorsal funiculus curves between a wildtype control group and two different mutant groups of mice. The homozygous control group had EphA4 completely conserved (genotype $\text{EphA4}^{\text{KI/KI}}$) and the heterozygous mutant mice had one wildtype allele of EphA4 and one mutant allele: one mutant mouse line was heterozygous (genotype $\text{EphA4}^{\text{KD/KI}}$) with the mutant allele harboring a point mutation in the tyrosine kinase domain located in the C-terminus of EphA4, which renders the receptor enzymatically inactive; the other mutant mouse line was heterozygous (genotype $\text{EphA4}^{\text{GFP/KI}}$) with the mutant allele having the complete C-terminus of the receptor knocked out and replaced by green fluorescent protein (GFP) (see Figure 5.7).

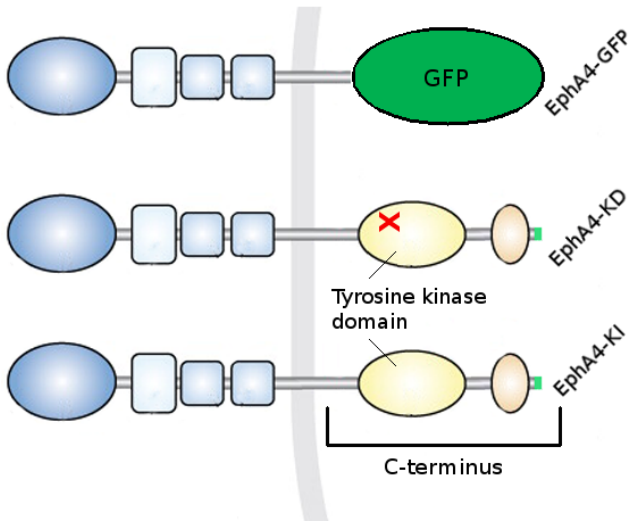


Figure 5.7: Schematic diagram showing the C-terminus of the EphA4 gene for the alleles KI (bottom), KD (medium), and GFP (top).

The standardized length of the dorsal funiculus was measured at 25 cross-sections along the spinal cord (see Figure 5.8) in five animals with genotype KI/KI, six animals with KD/KI, and four animals with GFP/KI. The measurements for all animals are displayed in Figure 5.1 at the beginning of this chapter.

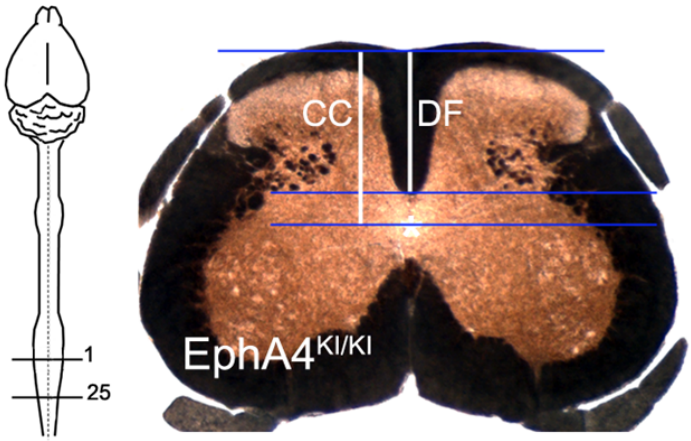


Figure 5.8: Range of the 25 cross-sections along the spinal cord (left figure) and cross-section with standardized length of the dorsal funiculus: DF/CC (right figure).

We modeled the curves of each group of animals by a semiparametric mixed model as described in Subsection 5.1.1:

$$y_{jik} = f_k(x_{jik}) + \alpha_{ik} + \varepsilon_{jik}, \tag{5.7}$$

with $\alpha_{ik} \sim N(0, \sigma_\alpha^2)$ and $\varepsilon_{jik} \sim N(0, \sigma_\varepsilon^2)$ for $K = 3$ groups $k = 1, \ldots, K$, animals $i = 1, \ldots, N(k)$ in the $k$th group, and $N(ik) = 25$ measurements $j = 1, \ldots, N(ik)$ for each animal. The number of animals $N(k)$ in the $k$th group are $N(1) = 5$, $N(2) = 5$, and $N(3) = 4$ with $k = 1$ corresponding to genotype KI/KI, $k = 2$ corresponding to genotype KD/KI, and $k = 3$ corresponding to genotype GFP/KI. This leads to $N = 375$ observations $y_{jik}$ in total. The smooth functions were approximated by a linear combination of B-splines basis functions (Eilers and Marx, 1996).

The fitted groupwise curves are shown in Figure 5.9. Tukey-type comparisons of the three curves were conducted, where each pair of curves was compared in each of the 25 positions. Significant differences with adjusted $p$ values $< 0.05$ were found on positions 1 to 20 when comparing the wildtype control to the mutant KD/KI, and on positions 1 to 9 for when comparing the wildtype control to the mutant GFP/KI. Hence, the kinase domain of the C-terminus is required for the development of the complete dorsal funiculus and one kinase-dead or kinase-absent allele already affects the dorsal funiculus compared to the wildtype with both alleles functioning normally. Absence of the C-terminus including the kinase domain (EphA4-GFP) leads to reduction of the dorsal funiculus in the lower positions, whereas inactivation of the kinase domain (EphA4-KD) leads to reduction of the dorsal funiculus in almost the entire region inspected. Significant differences with adjusted $p$ values $< 0.05$ were found in the medium region (positions 9-13) between the two heterozygous mutant mouse lines.

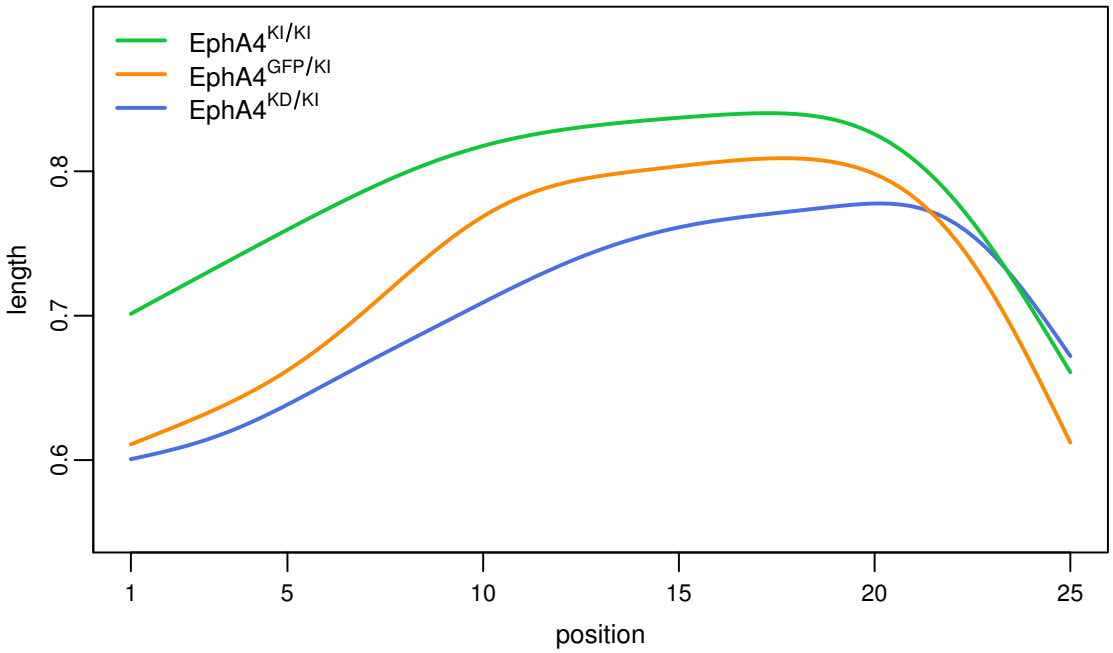


Figure 5.9: Curves of the length of the dorsal funiculus for the three genotypes estimated by model (5.7).

## 5.4 Behavior of Multiple AUC Comparisons

The following simulations aim to estimate the FWER and power of multiple comparisons of AUCs performed by the proposed approach in small samples and for different combinations of number of subjects per group and number of measurements per subject. A further objective is to investigate the influence of the measurement spacing on the properties by comparison of settings with equally spaced measurements for all subjects and random measurement points for each subject. The influence of the inevitable initial estimation of the AUC by the trapezoidal rule is estimated by comparing AUC estimates from a narrow grid to ones from a wider grid.

### 5.4.1 Simulation Setup

The FWER and power of the testing procedure for multiple comparisons of AUCs were estimated for Dunnett- and Tukey-type comparisons of the areas under three curves. We considered three functions $f_1$, $f_2$, and $f_3$ which share the same AUC on the interval $[0, 1]$ (see left part of Figure 5.10):

$$\begin{aligned} f_1(x) &= \frac{1}{3}\left(\sin(\pi(2x - 0.5)) + 2\right) \\ f_2(x) &= \frac{1}{3}\left(\sin(\pi(2x + 1)) + 2\right) \\ f_3(x) &= \frac{1}{3}\left(\sin(2\pi x) + 2\right) \end{aligned}$$

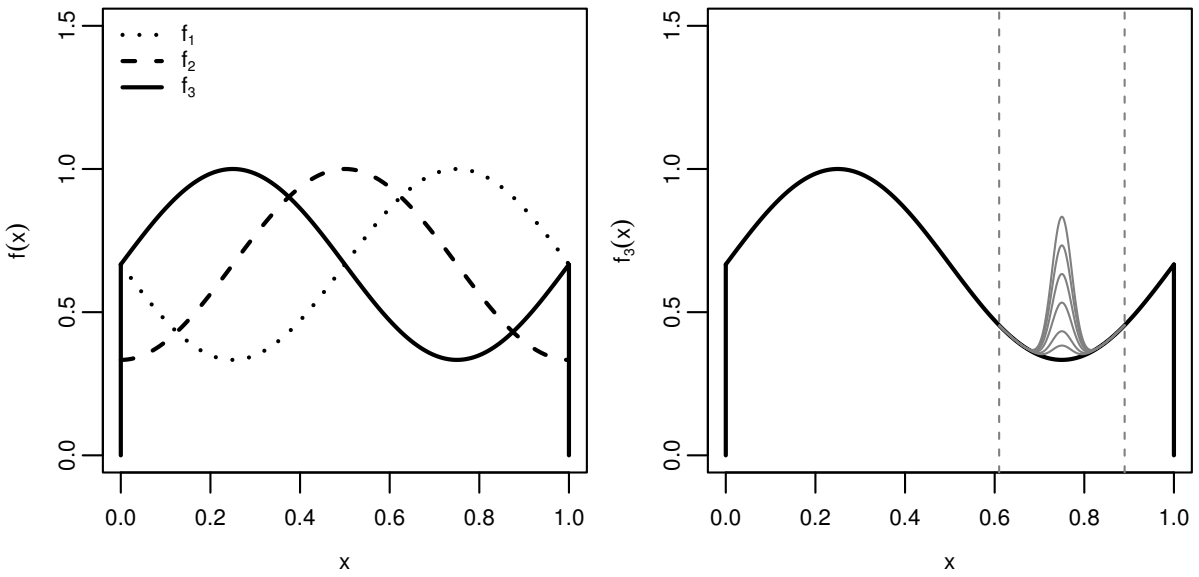


Figure 5.10: True smooth functions for the estimation of the familywise error rate for comparisons of AUC (left figure) and smooth function of group 3 (gray) for values of $a$ varying between 0 and 0.7 for the estimation of power (right figure).

The standard deviations for the error terms and the random effects were chosen in two different combinations:

- $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.02$,
- $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.05$,

Two different grid patterns for $x$ were considered:

(a) equally spaced on $[0, 1]$,

(b) continuous uniformly distributed on $[0, 1]$ with different positions for different subjects,

For $N(k)$ subjects in each group $k$, observations at $J(ik)$ values $x \in [0, 1]$ were simulated from the "true" function

$$y_{jik} = f_k(x_{jik}) + \alpha_{ik} + \varepsilon_{jik}, \tag{5.8}$$

with $\alpha_{ik} \sim N(0, 0.004)$ and $\varepsilon_{ijk} \sim N(0, 0.004)$. The same patterns of time points on $[0, 1]$ as described in Subsection 5.2.1 were used. We investigated scenarios with 20, 30, 40, or 50 time points and 5, 10, 15, or 20 subjects per group. For the estimation of the FWER, observations were simulated from the "null model" (5.8) for all three groups, i.e., the true functions differ between groups, but the areas under the curves are the same. The area under the curve was approximated from the fitted curves by the trapezoidal rule with as many grid points as measurements with equal spacing, which in setting (a) corresponds to

$$\int f(x)\, d\mathbb{P}_X,$$

and additionally for a more narrow grid, i.e.,

$$\int f(x)\,dx,$$

with the number of grid points chosen as 10 times the number of measurements equally spaced between the minimum and the maximum measurement point. The FWER was estimated by the portion of 1,000 datasets, in which at least one difference was found among all comparisons made, and the same datasets were used for both Dunnett- and Tukey-type comparisons. The nominal level was $\alpha = 0.05$.

To investigate the power of the procedure, the observations of group 3 were simulated from a function which differed from model (5.8) for $x$ values in the interval $[0.61, 0.89]$:

$$f_3^*(x) = f_3(x) + \exp\left(-\frac{(x-0.75)^2}{0.01}\right) \cdot a \cdot I_{[0.61,0.89]}(x), \tag{5.9}$$

$a \in [0, 0.7]$, leading to a higher AUC for group 3 compared to the other groups (see Figure 5.10). The power of the procedure was estimated by the portion of 1,000 simulated datasets, in which at least one true significant difference of areas was found (disjunctive power).

### 5.4.2 Simulation Results

#### Familywise Error Rate

The FWER for all simulation scenarios is shown in Figure 5.11.

The results are similar for Dunnett- and Tukey-type comparisons. The FWER is only slightly higher when measurement points were not equally spaced but at random positions for each subject (setting (b) versus setting (a)). No considerable difference can be detected between equal variances for errors and random effects and larger variances for errors (bottom row versus top row). For AUCs estimated from data sets with each group containing only 5 subjects, the procedure is liberal. The FWER reaches the nominal level when 20 subjects are inspected in each group.

#### Power

The estimated power curves for setting (a), 15 subjects per group, and 20, 30, 40 or 50 time points are shown in Figure 5.12 for Dunnett-type and Tukey-type comparisons.

There is no considerable increase in power with increasing number of measurements taken from each subject. The power curve ascends quickly over the inspected range of the parameter $a$. Rather small changes in the area under the curve of the third group are detected well.

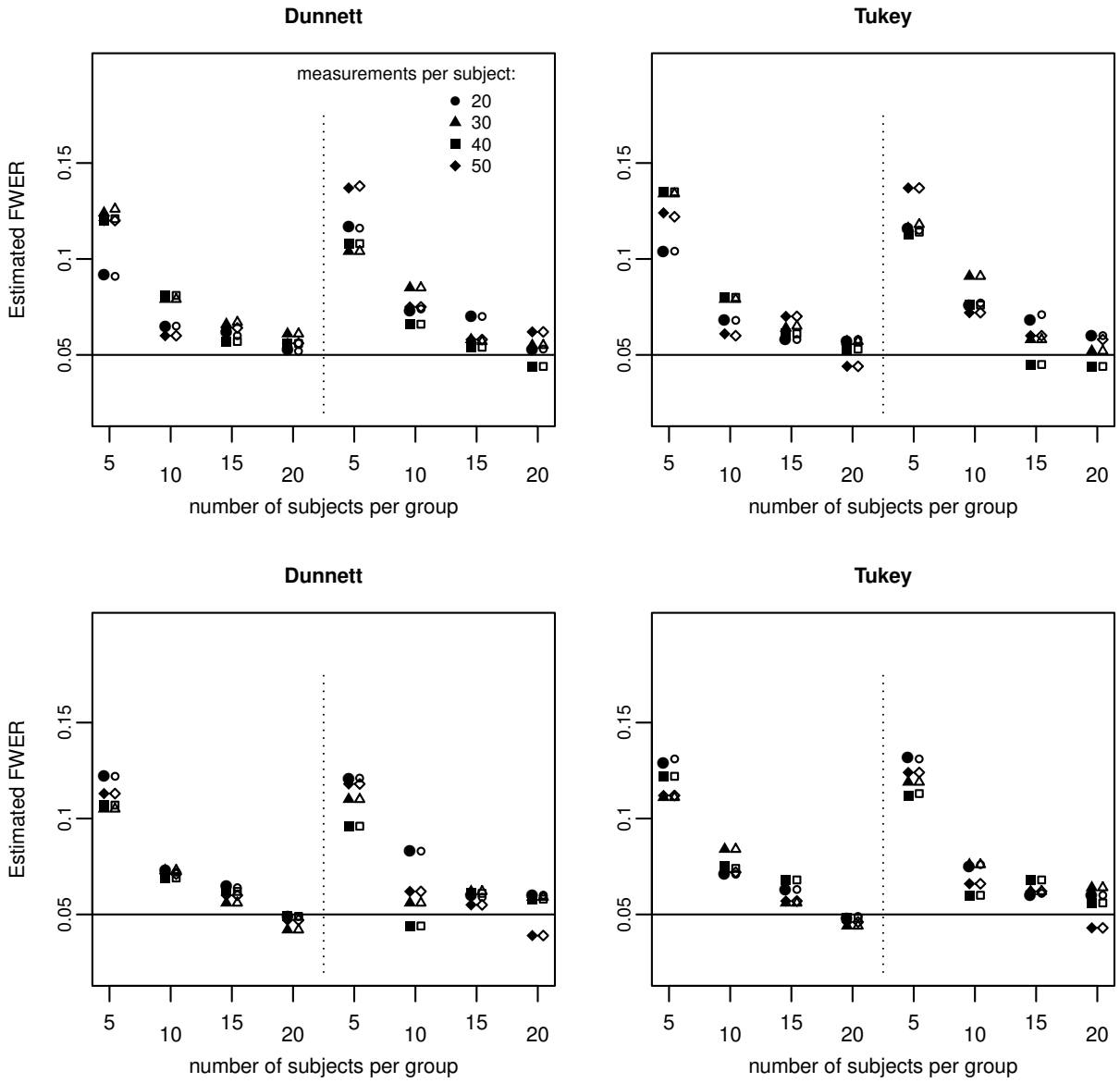


Figure 5.11: Estimated familywise error rate for Dunnett- and Tukey-type comparisons of three AUCs for equally spaced measurements (left section) and uniformly distributed measurements (right section) with standard deviations $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.02$ (upper row) and $\sigma_\alpha = 0.02$ and $\sigma_\varepsilon = 0.05$ (bottom row).

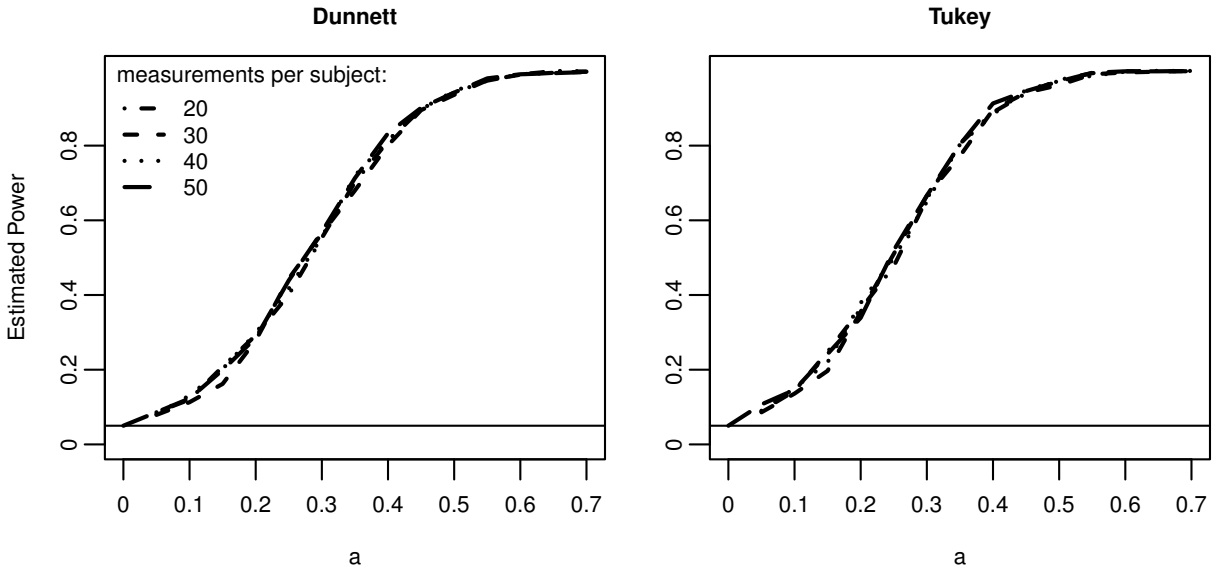


Figure 5.12: Power of Dunnett- and Tukey-type comparisons of three AUC where the AUC of one group differed from the others according to equation (5.8) for varying values of $a$.

## 5.5 Comparisons of Exposure Dosages of Benzene on Pre-Phenylmercapturic

Benzene is a compound used as a solvent in the chemical industry. In the 1980's concern arose whether chronic exposure to low levels of benzene has a toxic effect to humans (Sabourin, Bechtold, Birnbaum, Lucier, and Henderson, 1987). A study of the US National Institute of Environmental Health Sciences studied the effect of the exposure concentration on the benzene metabolism in mice and rats. Bailer (1988) presented data of this study comparing the post dose concentrations of pre-phenylmercapturic acid, a particular metabolite of benzene, between three exposure groups: 50 ppm for 6 hours, 150 ppm for 2 hours and 600 ppm for 0.5 hours. These data are reanalyzed using the approach for multiple comparisons of AUCs described earlier in this chapter. For each dose group four replicates were measured at five time points from 0 to 8 hours after the end of the treatment. The data are displayed in Figure 5.13.

The measurements followed a serial sampling design in which only one sample was taken from each mouse, because the mice were killed to obtain the metabolite concentration in the liver. Therefore, measurements can be assumed to be independent and group-level concentration-time curves can be fit using the model

$$y_{jik} = f_k(x_j) + \varepsilon_{jik}, \tag{5.10}$$

with $y_{jik}$ the concentration of the $i$th mouse in group $k$ taken at the $j$th time point $x_j$ and each mouse measured only at one time point. $\varepsilon_{jik} \sim N(0, \sigma_\varepsilon^2)$ are independent normal

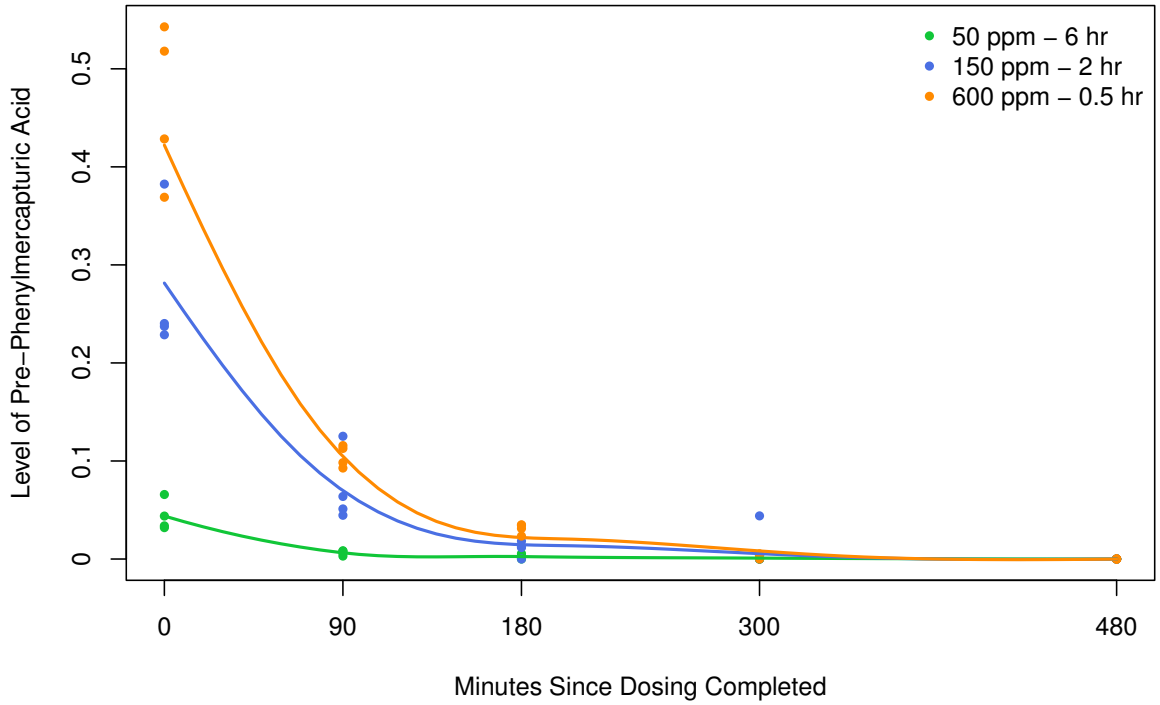


Figure 5.13: Measurements of the pre-phenylmercapturic acid level ($\mu$mole/g) at several times post dose in the livers of mice exposed to different dose rates and group-level curves estimated by model (5.10).

random errors for each measurement. The model corresponds to model (5.1) without the random subject-specific effect. Figure 5.13 shows the estimated group-level curves, which are estimated by approximating each curve by a penalized spline and employing BLUP estimation in the corresponding linear mixed model representation. AUCs can be estimated by linear combinations of the estimated model parameters using the trapezoidal rule on a narrow grid with 50 nodes. Hypotheses are formulated as contrasts such that each contrast compares the AUC between two exposure groups. Adjusted $p$-values for each of the three AUC comparisons were $< 10^{-6}$, i.e., the body exposure to drug differs pairwise between all three exposure groups. This result agrees with the result in Bailer (1988).

## 5.6 Summary

In this chapter, a method for pairwise comparisons of several populations of curves fitted by a semiparametric mixed model and for multiple comparisons of the curves thereunder was derived. The approach is based on the simultaneous inference procedure by Hothorn et al. (2008). Other methods for comparisons of several groups only allow for overall comparisons and do not provide information on the time points or positions at which the curves differ between two groups. The method for multiple curve comparisons proposed in this chapter allows for pairwise comparisons of two or more groups over a grid along

the curves with control of the probability of at least one false-positive finding among all comparisons made. A FWER close to the nominal level is achieved in settings with reasonable sample size. However, the power of the procedure to detect a true difference between curves is rather low in the scenario inspected.
One of the benefits of the method is information about the sections in which the curves differ. Comparisons of the dorsal funiculus curves between groups of mice with different EphA4 genotypes using the new method gave information about which region of the spinal cord is sensitive to lack of certain EphA4 domains.
An extension of the method can be used for comparisons of the areas under multiple curves. This is relevant e.g. to multiple comparisons of areas under concentration-time curves in pharmacokinetic studies with the objective of comparing body exposure to drug between different exposure groups. In contrast to the approach by Bailer (1988) the estimation of the AUC is based on a flexible estimation of the concentration-time curve and not on sample means at each time point. If measurements over time are taken on the same subject, the resulting correlation can be modeled by a random effect in the semiparametric mixed model used to estimate the curves. Simulation showed a FWER close to the nominal level for sufficiently large samples and good power properties. However, the sample sizes commonly used in pharmacokinetic studies are not large enough to ensure proper FWER control.

# Chapter 6

# Conclusion

Nonstandard parametric designs are frequently needed to describe the complex data structure arising in applied research and multiple comparisons are required within these designs to address the study objectives. The framework for simultaneous inference in general parametric models by Hothorn et al. (2008) is a flexible tool for multiple comparisons in parametric models with generally correlated parameter estimates. By adequate specification of the contrast matrix it can be applied to a variety of research questions. However, due to the asymptotic nature of the reference distribution the procedure controls the error rate across all comparisons only for sufficiently large samples.

This thesis evaluated the small samples properties of the simultaneous inference procedure embedded in complex parametric designs, which reflect settings from medical and biological research. In summary, the familywise error rate is controlled well for sample sizes commonly available in these studies and the power is sufficient to detect relevant differences.

Investigating differences between means of more than two groups is often a scientific hypothesis under test. Numerous procedures for overall comparisons and pairwise comparisons exist but these use the standard assumption of homogeneous variances among groups. For these classical procedures the presence of heteroscedasticity leads to a deviation of the familywise error from the nominal level $\alpha$ which persists with increasing sample size. By using a heteroscedastic-consistent covariance estimation technique, the method can be used for multiple comparisons in presence of either equal or unequal group variances in balanced or unbalanced designs with arbitrary error distribution. Simulations showed that the familywise error rate is bound by the nominal level already for relatively small sample sizes in unbalanced designs with both normal or skewed error distributions and different kinds of pairing of group sizes and variance, whereas the Tukey-Kramer test can lead to false positive rates considerably higher than $\alpha$. Even in situations where the Tukey-Kramer test does not lead to inflated false positive rates, simultaneous inference employing heteroscedastic-consistent covariance estimation is superior to the Tukey-Kramer test, because it has the greater power. The approach for simultaneous inference leads to valid conclusions about differences between sample means in presence of unequal

variances and can be recommended especially when there are doubts over homoscedasticity.

Multiple comparisons with a control are a common issue in multicenter clinical trials with survival endpoint. The simultaneous inference procedure proposed by Hothorn et al. (2008) can be applied to many-to-one comparisons in the frailty Cox model. The method allows clustered survival data of several experimental treatments to be compared overall and pairwise with a control with adjustment for covariates. Limits for one-sided simultaneous confidence intervals for hazard ratios, which compare several experimental treatments each with a control, can be obtained using penalized likelihood estimation of the parameter estimates and then employing the approach for simultaneous inference in parametric models. Compared to multiplicity-adjusted $p$-values, the simultaneous confidence intervals for hazard ratios can be interpreted in terms of clinical importance. Covariate information can be included, and the multicenter structure can be taken into account. Simulations showed that the joint coverage probability for many-to-one comparisons of treatment effects is close to the nominal confidence level of $1 - \alpha$ even for relatively small sample sizes, both in balanced and in unbalanced designs. Especially in the settings with a large number of clusters with few observations within each cluster, which often occurs in multicenter clinical trials, the coverage probability of the simultaneous confidence intervals equaled the nominal level. The joint coverage probability of the simultaneous confidence intervals decreases systematically with increasing distance of the effect size from the null hypothesis caused by bias in the parameter estimates. However, for the values of hazard ratios usually reported in randomized trials, the bias of the estimates is relatively small and the joint coverage of the confidence intervals very close to $1 - \alpha$ for samples of reasonable size. The method is superior to the current approach, which uses the Bonferroni correction in the Cox model to adjust for multiple comparisons of treatments and suffers from low power.

Simultaneous inference about Williams-type contrasts for the comparison of survival between several dose groups and a control group was evaluated in survival models. The procedure can be used to analyze the toxicological endpoint survival by a Williams-type procedure, analogous to the National Toxicology Program practice for normal endpoints. In the case of proportional hazards between treatment groups time of survival can be modeled using the Cox proportional hazards model. In the presence of non-proportional hazards the accelerated failure time model should be used instead. Difference of survival between two groups can be measured by the hazard ratio for the Cox PH model and by the survival time ratio for the AFT model respectively. The simultaneous confidence intervals for the hazard/survival time ratios as the effect size allow interpretation according to both statistical significance and biological relevance. Simulations indicate both the control of the familywise error and appropriate power for the balanced design with sample sizes according to the recommendations of the National Toxicology Program. The approach is powerful to detect different dose response shapes and can be recommended for the analysis of mortality in toxicological studies.

A further part of this thesis addressed pairwise comparisons of several populations of

curves fitted by a semiparametric mixed model. Previously existing approaches only perform overall comparisons and do not provide information on the time points or positions in which the curves differ between two groups. The presented approach allows for pairwise comparisons of two or more groups over a grid along the curves with control of the probability of at least one false-positive finding among all comparisons made. Simulations show control of an overall error level for multiple comparisons of several curves fitted from a reasonable number of observations per subject and per group when curves were compared in the positions in which the measurements were taken.

The methodology can be extended to multiple comparisons of areas under curves in settings where a parametric model for the curves and an estimate of the areas thereunder does not exist. By using a semiparametric mixed model to fit the curves and applying the trapezoidal rule to estimate the area under the curves the procedure by Hothorn et al. (2008) can be applied to perform multiple comparisons of the areas under the curves. However, in this multiple testing problem control of the familywise error rate is only achieved in samples with considerably more observations than commonly available in studies which address comparisons of areas under the curves. Whether alternative procedures suffer from the same problem needs to be addressed by further research.

Several applications illustrated the necessity for multiple comparison procedures in designs beyond the standard models, which can be addressed by the simultaneous inference procedure for general parametric models. The simulation studies in this thesis deliver the satisfactory result that the probability of any false positive finding can be controlled already in samples of small size. The approach can thus be recommended for multiple comparisons in applied research without hesitation.

# Bibliography


Bailer, A. J. (1988). Testing for the equality of area under the curves when using destructive measurement techniques. *Journal of Pharmacokinetics and Pharmacodynamics*, **16**(3), 303–309.

Barker, P., and Henderson, R. (2005). Small sample bias in the gamma frailty model for univariate survival. *Lifetime Data Analysis*, **11**(2), 265–284.

Behseta, S., and Chenouri, S. (2011). Comparison of two populations of curves with an application in neuronal data analysis. *Statistics in Medicine*, **30**(12), 1441–1454.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, **24**(11), 1713–1723.

Berger, V. W. (2004). On the generation and ownership of alpha in medical studies. *Controlled Clinical Trials*, **25**(6), 613–619.

Bertrand, J., Comets, E., Chenel, M., and Metré, F. (2012). Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. *Biometrics*, **68**(1), 146–155.

Bhadra, D., Daniels, M. J., Kim, S., Ghosh, M., and Mukherjee, B. (2012). A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case-control studies. *Biometrics*, **68**(2), 361–370.

Bolin, R. W., Robinson, W. A., Sutherland, J., and Hamman, R. F. (1982). Busulfan versus hydroxyurea in long-term therapy of chronic myelogenous leukemia. *Cancer*, **50**(11), 1683–1686.

Bretz, F. (1999). *Powerful Modifications of Williams' Test on Trend*. Ph.D. thesis, Institut für Biostatistik der Universität Hannover.

Bretz, F., Genz, A., and Hothorn, L. A. (2001). On the numerical availability of multiple comparison procedures. *Biometrical Journal*, **43**(5), 645–656.

Bretz, F., and Hothorn, L. A. (2000). A powerful alternative to Williams' test with application to toxicological dose-response relationship of normally distributed data. *Environmental and Ecological Statistics*, **7**(2), 135–154.

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

Bretz, F., Hothorn, T., and Westfall, P. (2010). *Multiple Comparisons Using R*. Boca Raton, Florida, USA: Chapman & Hall/CRC Press.

Brown, M. B., and Forsythe, A. B. (1974a). The ANOVA and multiple comparisons for data with heterogenous variances. *Biometrics*, **30**(4), 719–724.

Brown, M. B., and Forsythe, A. B. (1974b). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, **16**(1), 129–132.

Butts, C. A., Ding, K., Seymour, L., Twumasi-Ankrah, P., Graham, B., Gandara, D., Johnson, D. H., Kesler, K. A., Green, M., Vincent, M., Cormier, Y., Goss, G., Findlay, B., Johnston, M., Tsao, M.-S., and Shepherd, F. A. (2010). Randomized phase III trial of vinorelbine plus cisplatin compared with observation in completely resected stage IB and II non–small-cell lung cancer: updated survival analysis of JBR-10. *Journal of Clinical Oncology*, **28**(1), 29–34.

Chakraborti, S., and Desu, M. M. (1991). Linear rank tests for comparing treatments with a control when data are subject to unequal patterns of censorship. *Statistica Neerlandica*, **45**(3), 227–254.

Chen, Y. I. (2000). Multiple comparisons in carcinogenesis study with right-censored survival data. *Statistics in Medicine*, **19**(3), 353–367.

Chuang-Stein, C., and Tong, D. M. (1995). Multiple comparisons procedures for comparing several treatments with a control based on binary data. *Statistics in Medicine*, **14**(23), 2509–2522.

Cohan, F., and Perry, E. B. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Current Biology*, **17**(10), 373–386.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **34**(2), 187–220.

Cribato-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, **45**(2), 215–233.

Cribato-Neto, F., and da Silva, W. B. (2011). A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. *Advances in Statistical Analysis*, **95**(2), 129–146.

Cribato-Neto, F., Souza, T. C., and Vasconcellos, K. L. P. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics – Theory and Methods*, **36**(10), 1877–1888.

Davidson, R., and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**(272), 1096–1121.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, **75**(372), 796–800.

Dunnett, C. W., and Tamhane, A. C. (1992). Comparisons between a new drug and active and placebo controls in an efficacy clinical trial. *Statistics in Medicine*, **11**(8), 1057–1063.

Egea, J., and Klein, R. (2007). Bidirectional Eph-ephrin signaling during axon guidance. *Trends in Cell Biology*, **17**(5), 230–238.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimator for families of linear regression. *The Annals of Mathematical Statistics*, **34**(2), 447–456.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89–121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, **14**(3), 715–745.

Games, P. A., and Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, **1**(2), 113–125.

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

Genz, A., and Bretz, F. (2002). Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics*, **11**(4), 950–971.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, **87**(420), 942–951.

Hasford, J., Pfirrmann, M., Hehlmann, R., Allan, N. C., Baccarani, M., Kluin-Nelemans, J. C., Alimena, G., Steegmann, J. L., and Ansari, H. (1998). A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. Writing committee for the collaborative CML prognostic factors project group. *Journal of the National Cancer Institute*, **90**(11), 850–858.

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

Hehlmann, R., Heimpel, H., Hasford, J., Kolb, H. J., Pralle, H., Hossfeld, D. K., Queisser, W., Löffler, H., Hochhaus, A., and Heinze, B. (1994). Randomized comparison of interferon-alpha with busulfan and hydroxyurea in chronic myelogenous leukemia. The German CML Study Group. *Blood*, **84**(12), 4064–4077.

Herberich, E. (2009). *Niveau und Güte simultaner parametrischer Inferenzverfahren*. Diploma thesis, Ludwig-Maximilians-Universität München.

Herberich, E., and Hothorn, L. A. (2012a). Statistical evaluation of mortality in long-term carcinogenicity bioassays using a Williams-type procedure. *Regulatory Toxicology and Pharmacology*, **64**(1), 26–34.

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

Herberich, E., Sikorski, J., and Hothorn, T. (2010). A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PLoS ONE*, **5**, e9788.

Hoaglin, D. C., and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, **32**(1), 17–22.

Hochberg, E., and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: John Wiley & Sons.

Hollander, M., and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, 2nd Ed.

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

Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. (1982). Covariate analysis of survival data: A small-sample study of Cox's model. *Biometrics*, **38**(3), 685–698.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, 2nd Ed.

Klingenberg, B. (2010). Simultaneous confidence bounds for relative risks in multiple comparisons to control. *Statistics in Medicine*, **29**(30), 3232–3244.

Koeppel, A., Perry, E. B., Sikorski, J., Krizanc, D., Warner, A., and Ward, D. M. (2008). Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(7), 2504–2509.

Kong, M., and Yan, J. (2011). Modeling and testing treated tumor growth using cubic smoothing splines. *Biometrical Journal*, **53**(4), 1–19.

Konietschke, F. (2009). *Simultane Konfidenzintervalle für nichtparametrische relative Kontrasteffekte*. Ph.D. thesis, Georg-August-Universität Göttingen.

Krafty, R. T., Gimotty, P. A., Holtz, D., Coukos, G., and Guo, W. (2008). Varying coefficient model with unknown within-subject covariance for analysis of tumor growth curves. *Biometrics*, **64**(4), 1023–1031.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, **12**(3), 307–310.

Kulathinal, S., and Gasbarra, D. (2002). Testing equality of cause-specific hazard rates corresponding to m competing risks among K groups. *Lifetime Data Analysis*, **8**(2), 147–161.

Kullander, K., Mather, N. K., Diella, F., Dottori, M., Boyd, A. W., and Klein, R. (2001). Kinase-dependent and kinase-independent functions of EphA4 receptors in major axon tract formation in vivo. *Neuron*, **29**(1), 73–84.

Lee, S., and Ahn, C. (2003). Modified ANOVA for unequal variances. *Communications in Statistics – Simulation and Computation*, **32**(4), 987–1004.

Logan, B. R., Wang, H., and Zhang, M. J. (2005). Pairwise multiple comparison adjustment in survival analysis. *Statistics in Medicine*, **24**(16), 2509–2523.

Long, J. S., and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, **54**(3), 217–224.

MacKinnon, J. G., and White, H. (1985). Some heteroscedasticity consistent covariance with improved finite sample properties. *Journal of Econometrics*, **29**(3), 305–325.

Moore, M. J., Goldstein, D., Hamm, J., Figer, A., Hecht, J. R., Gallinger, S., Au, H. J., Murawa, P., Walde, D., Wolff, R. A., Campos, D., Lim, R., Ding, K., Clark, G., Voskoglou-Nomikos, T., Ptasynski, M., and Parulekar, W. (2007). Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *Journal of Clinical Oncology*, **25**(15), 1960–1966.

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

National Toxicology Program (U.S.) (1979). Bioassay of piperonyl butoxide for possible carcinogenicity (CAS no. 51-06-6 / NCI-CG-TR-120). Technical report.

National Toxicology Program (U.S.) (2012). Testing information – Descriptions of NTP study types – Statistical procedures – Expanded overview. Available from <http://ntp.niehs.nih.gov> (accessed on 26 Aug 2012).

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *The Annals of Statistics*, **26**(1), 183–214.

Poplin, E., Feng, Y., Berlin, J., Rothenberg, M. L., Hochster, H., Mitchell, E., Alberts, S., O'Dwyer, P., Haller, D., Catalano, P., Cella, D., and Benson, A. B. (2009). Phase III, randomized study of gemcitabine and oxaliplatin versus gemcitabine (fixed-dose rate infusion) compared with gemcitabine (30-minute infusion) in patients with pancreatic carcinoma E6201: A trial of the Eastern Cooperative Oncology Group. *Journal of Clinical Oncology*, **27**(23), 3778–3785.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL `http://www.R-project.org/`

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Sabourin, P. J., Bechtold, W., Birnbaum, L., Lucier, G., and Henderson, R. (1987). Water-soluble benzene metabolites in F344/N rats and B6C3F$_1$ mice during and following $^3$H-benzene inhalation. *Toxicologist*, **7**, 232.

Saltz, L. B., Cox, J. V., Lanke, C. B., Rosen, L. S., Fehrenbacher, L., Moore, M. J., Maroun, J. A., Ackland, S. P., Locker, P. K., Pirotta, N., Elfring, G. L., and Miller, L. L. (2001). Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. *New England Journal of Medicine*, **343**(13), 905–914.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, **2**(6), 110–114.

Schaarschmidt, F., Biesheuvel, E., and Hothorn, L. A. (2009). Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. *Journal of Biopharmaceutical Statistics*, **19**(2), 292–310.

Senn, S., and Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, **6**(3), 161–170.

Serfling, R. J. (1980). *Approximation Theorems for Mathematical Statistics*. New York: John Wiley & Sons.

Sikorski, J., Brambilla, E., Kroppenstedt, R. M., and Tindall, B. J. (2008a). The temperature adaptive fatty acid content in Bacillus simplex strains from 'Evolution Canyon', Israel. *Microbiology*, **154**(8), 2416–2426.

Sikorski, J., and Nevo, E. (2005). Adaptation and incipient sympatric speciation of Bacillus simplex under microclimatic contrast at "Evolution Canyons" I and II, Israel. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(44), 15924–15929.

Sikorski, J., Pukall, R., and Stackebrandt, E. (2008b). Carbon source utilization patterns of Bacillus simplex ecotypes do not reflect their adaptation to ecologically divergent slopes in 'Evolution Canyon', Israel. *FEMS Microbiology Ecology*, **66**(1), 38–44.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **47**(1), 1–52.

Tamhane, A. C. (1977). Multiple comparisons in model I one-way ANOVA with unequal variances. *Communications in Statistics – Theory and Methods*, **6**(1), 15–32.

Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, **74**(366a), 471–480.

Tarone, R. E. (1975). Tests for trend in life table analysis. *Biometrika*, **62**(3), 679–690.

Therneau, T. (2012a). *A Package for Survival Analysis in S*. R package version 2.36-14.

Therneau, T. M. (2012b). coxme: Mixed effects cox models. R package version 2.2-3. URL `http://CRAN.R-project.org/package=coxme`

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

Treat, J. A., Gonin, R., Socinski, M. A., Edelman, M. J., Catalano, R. B., Marinucci, D. M., Ansari, R., Gillenwater, H. H., Rowland, K. M., Comis, R. L., Obasaju, C. K., and Belani, C. P. (2010). A randomized, phase III multicenter trial of gemcitabine in combination with carboplatin or paclitaxel versus paclitaxel plus carboplatin in patients with advanced or metastatic non-small-cell lung cancer. *Annals of Oncology*, **21**(3), 540–547.

Tukey, J. W. (1953). The problem of multiple comparisons. Dittoed manuscript of 386 pages, New Jersey: Department of Statistics, Princeton University.

Weerahandi, S. (1995). ANOVA under unequal variances. *Biometrics*, **51**(2), 589–599.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, **38**(3–4), 330–336.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute Inc.

Westlake, W. (1973). Use of statistical methods in evaluation of in vivo performance of dosage forms. *Journal of Pharmaceutical Sciences*, **62**(10), 1579–1589.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test of heteroskedasticity. *Econometrica*, **48**(4), 817–838.

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

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

Wood, S. (2006a). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, S. (2006b). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, **48**(4), 445–464.

Xu, L., and Wang, S. (2008). A new generalized $p$-value and its upper bound for ANOVA under unequal error variances. *Communications in Statistics – Theory and Methods*, **37**(7), 1002–1010.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, **11**(10), 1–17.
URL `http://www.jstatsoft.org/v11/i10/`

Zhang, D., Lin, X., and Sowers, M. (2000). Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. *Biometrics*, **56**(1), 31–39.

## Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Herberich, Esther
Name, Vorname

Ort, Datum

Unterschrift Doktorand/in