Evaluation of interaction effects in two-factorial designs by simultaneous confidence intervals in the cell means model

Von der Naturwissenschaftlichen Fakultät
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades

Doktor der Naturwissenschaften
– Dr.rer.nat. –
genehmigte Dissertation
von

M.Sc. **Andreas Kitsche**

geboren am 10.02.1986 in Zerbst

2014

Referent: Prof. Dr. Ludwig A. Hothorn
Korreferent: Prof. Dr. Hans-Peter Piepho
Tag der Promotion: 06.02.2014

Kurzzusammenfassung

In den Biowissenschaften werden häufig Experimente mit zwei oder mehr Einflussfaktoren durchgeführt. Im Gegensatz zu einfaktoriellen Versuchen ermöglichen solche Anlagen sowohl die Analyse der Haupteffekte der einzelnen Einflussfaktoren, als auch einer etwaigen Interaktion zwischen den Faktoren. Dabei bezeichnet eine Interaktion im statistischen Sinne die unterschiedliche Ausprägung des Effektes eines Faktors über die Stufen eines anderen Faktors. In der klassischen Datenanalyse werden solche Interaktionen mittels einer Varianzanalyse und den korrespondierenden F-Tests ausgewertet. Ein Nachteil dieser Auswertung ist, dass der F-Test nur eine globale Testentscheidung zur Verfügung stellt. Häufig ist der Wissenschaftler jedoch daran interessiert auf welchen Faktorstufen die Interaktion beruht.

In dieser Arbeit wird die Formulierung von geeigneten Hypothesen für eine tiefgründigere Analyse der Interaktionseffekte dargestellt. Dabei werden die Hypothesen als Kontraste von Mittelwertsdifferenzen formuliert. Eine weitere Möglichkeit geeignete Hypothesen für Interaktionen auszudrücken besteht in der Bildung des Quotienten von Mittelwertsunterschieden. Diese Methode erlaubt es zusätzlich zwischen quantitativen und qualitativen Interaktionen zu unterscheiden.

Da bei den genannten Verfahren mehrere Hypothesen simultan untersucht werden, muss ein entsprechendes multiples Testverfahren angewendet werden, welches die Kontrolle des versuchsbezogenen Fehlers 1. Art erlaubt. Neben multiplizitäts-adjustierten p-Werten bieten simultane Konfidenzintervalle für die Interaktionseffekte die Möglichkeit die Größe und die Richtung der Effekte zu quantifizieren.

Die verwendeten Verfahren sind für die Anwendung auf Daten geeignet, bei denen angenommen wird, dass die primäre Zielvariable normalverteilt ist. Dabei wird sowohl der Fall unter der Annahme homogener Varianzen über die Gruppen, als auch unter der Annahme der Varianzheterogenität betrachtet. Weiterhin wird die Vorgehensweise für dichotome Endpunkte dargestellt.

Die vorgestellten Methoden zur Interaktionsanalyse werden anhand von Realdatenbeispielen aus biomedizinischen und gartenbaulichen Versuchen angewendet. Der entsprechende R–Code für die Auswertung wird kommentiert zur Verfügung gestellt.

Schlagworte: Multiple Kontrasttests, Simultane Konfidenzintervalle, Qualitative Interaktion

Abstract

Two or higher order factorial designs are very common in biomedical, agricultural, and horticultural research. Those factorial trials are appropriate to investigate beside the main effect of the factors also the interaction between them. The interaction effect determines the different response of one factor over the levels of the other factor. The commonly used evaluation of those trials by using the analysis of variance and the corresponding F-tests for the interaction effects offers only a global decision concerning the presence of interactions.

Within this thesis a straightforward method for the construction of appropriate hypotheses for an in depth analysis of statistical interactions is presented. The hypotheses are formulated via contrasts of differences among means. As an alternative approach, the hypotheses to test for interactions can be formulated as ratios of differences among means. In addition, this procedure allows the distinction between quantitative and qualitative interactions.

Because within such a detailed analysis several hypotheses are tested simultaneously, an adequate multiple comparison procedure has to be used. Multiplicity adjusted p-values for the individual hypotheses are provided, such that the significance can be inferred, while controlling the overall probability of a type I error. Furthermore, compatible simultaneous confidence intervals can be used to interpret the direction, magnitude and the biological relevance of the interaction effects.

The proposed methods are applicable on data with a normally distributed endpoint, whereas the cases of homogeneous and heterogeneous variances over the groups are considered. Additionally, the methodology is extended for binary response variables. The proposed methods are applied to two horticultural and three biomedical trials. In addition, the corresponding R code with comments for the reproducible analysis is provided.

In summary the main improvements of the presented methodology are: (i) providing inferences on global and local hypotheses (ii) evaluation of reasonable research hypotheses by user defined contrasts (iii) identification of the source of the interactions (iv) quantification of the interaction on a percentage scale (v) evaluation of the biological relevance of a potential interaction effect (vi) extension though the assessment of non-inferiority of the treatment effects.

Keywords: Multiple Contrast Tests, Simultaneous Confidence Intervals, Qualitative Interaction

Contents

List of Tables

List of Figures

Chapter 1.

Introduction

Experiments that include two or more factors are frequently conducted in a wide variety of research areas, such as horticultural, agricultural, biological and biomedical sciences. Those factorial trials permit the experimenter to simultaneously investigate several factors within the same experiment. As opposed to single-factor experiments, two and higher order factorial experiments are appropriate to investigate the interaction effects between the factors beside the main effects of each factor. According to Searle [1997] the interaction effect describes the extent to which one factor is not acting in the same manner over the levels of another factor. Depending on the research area the term interaction has also different meanings, which frequently leads to some confusion. Wang et al. [2010] provide an overview of several alternative meanings for the term interaction, especially differentiating the statistical and biological aspects. Within this thesis the term interaction refers to statistical interaction in factorial designs: How does the effect of one factor changes, if a second factor is varied. Furthermore, the focus lies on the analysis of first-order interactions, i.e. the interaction between two factors. Nevertheless, the presented methods can be extended to designs with more than two factors whereas the interpretation of higher-order interactions becomes more difficult.

A further problem on the analysis of statistical interactions is the distinction between quantitative and qualitative interactions, see Peto [1982] (cited by Gail and Simon [1985]). A quantitative interaction occurs if the effect of one factor varies in its magnitude but not in its sign over the levels of another factor. In contrast, a qualitative interaction denotes a difference in the effect of one factor

in its magnitude and in its sign over the levels of another factor. Figure 1.1 displays the different types of interactions for a two-by-two factorial design. It should be pointed out that a qualitative interaction due to the first factor does not automatically imply a qualitative interaction due to the second factor (see Figure 1.1 d and e). Since a difference of the effect of one factor over the levels of the other is expected in some research fields, a potential heterogeneity caused by a qualitative interaction is of particular interest [Gail and Simon, 1985]. Depending on the context under consideration several synonymous terms are used to denote qualitative and quantitative interactions, e.g., removable interaction or ordinal interaction, and non-removable or disordinal interactions [Gonzalez and Cox, 2007]. In agricultural experiments the term crossover interaction is used synonymously to qualitative interactions [Baker, 1988]. When analysing genotype-by-environmental interactions researcher are particularly interested in crossover interactions, which denote rank changes between environments within genotypes [Truberg and Hühn, 2000].

Besides the analysis of genotype-by-environmental interaction, the analysis of interaction between two treatment factors is of interest in horticultural and agricultural research. Some examples of treatment factors are: (i) particular varieties or cultivars of species [Compton, 2000], (ii) different kinds of fertilizers [Sahin et al., 2012], (iii) different irrigation intensities [Slauenwhite and Qaderi, 2013] or (iv) different seed spacing in a row [Petersen, 1985].

In most two-factorial biomedical experiments the factors can be classified in a primary treatment factor and a secondary factor [Cox, 1984]. According to the type of the secondary factor the following sources of interactions between two factors are relevant in biomedical research: (i) treatment-by-centre interaction in multi-centre trials, in which the secondary factor refers to medical centres or clinics [Potthoff et al., 2001], (ii) inconsistency in multi-regional trials, where the subjects are incorporated from many countries or regions [Chen et al., 2010], (iii) subgroup heterogeneity in trials with pre-defined subgroups, where the subsets of patients are defined by prognostic factors, e.g., age or disease severity [Gail and Simon, 1985], (iv) genome-wide association studies with focus on gene-gene and gene-environment interactions, in which the subsets are defined by a potential environmental factor such as smoking status [Han et al., 2012], (v) biomarker-by-

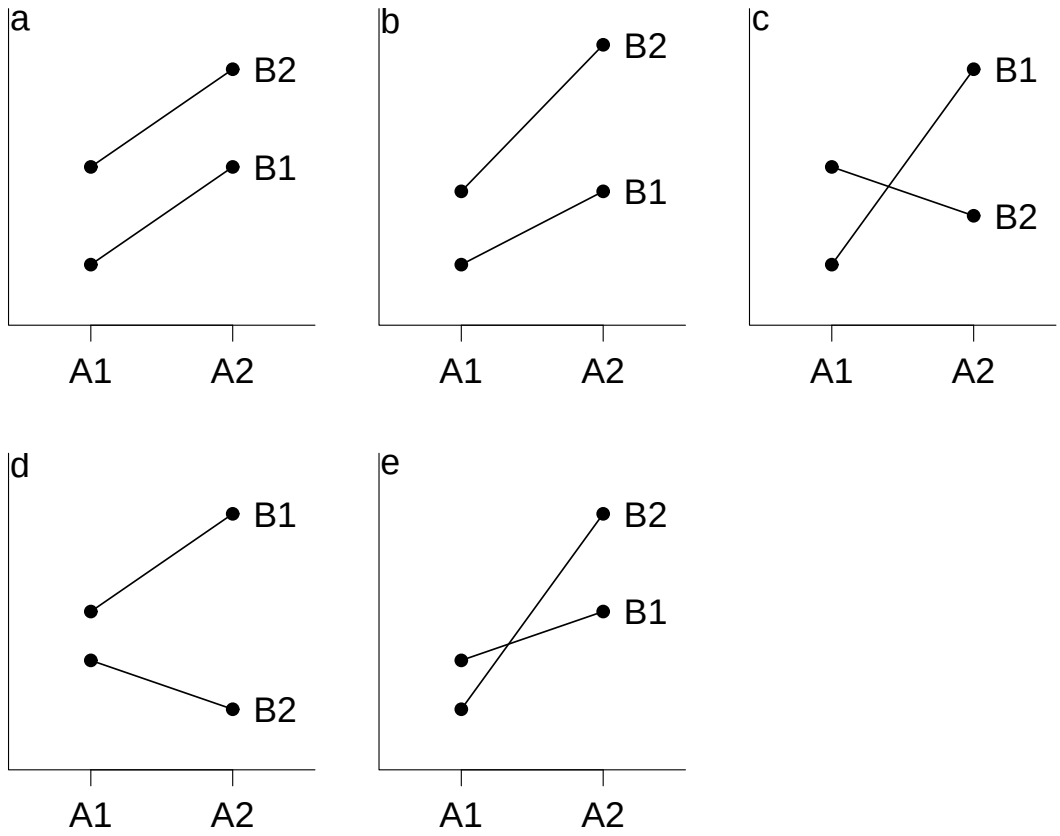

Figure 1.1.: Schematic display of different types of interactions for a two-by-two factorial design with the two factors A (with its levels $A1$ and $A2$) and B (with its levels $B1$ and $B2$). **(a)** no interaction: the difference between the levels of factor A is equal over the levels of factor B and vice versa, **(b)** quantitative interaction: the difference between the levels of factor A differs in its magnitude but not in its sign over the levels of factor B and vice versa **(c)** qualitative interaction: the difference between the levels of factor A differs in its magnitude and in its sign over the levels of factor B and vice versa **(d)** qualitative interaction according to factor A and quantitative interaction according to factor B, **(e)** qualitative interaction according to factor B and quantitative interaction according to factor A.

treatment interactions [Michiels et al., 2011], and (vi) heterogeneity between the different stages of adaptive trials [Parker, 2010].

Several regulatory guidelines address the analysis of interactions in multi-centre and multi-regional trials. Among others, the Guideline on Statistical Principles for Clinical Trials proposed by the ICH E9 [1998] refers to procedures when treatment-by-centre interactions are to be analysed: "*Marked heterogeneity may be identified by graphical display of the results of individual centres or by analytical methods, such as a significance test of the treatment-by-centre interaction*". In the answer of question 11 of the Question and Answers document ICH E5 [2009] the objectives of a multi-regional study for the purpose of bridging are defined: "*(1) to show that the drug is effective in the region and (2) to compare the results of the study between the regions with the intent of establishing that the drug is not sensitive to ethnic factors.*", whereas "not sensitive" means a consistent treatment effect over regions. However no recommendations are given on the test to use and how to assess a true heterogeneity of the treatment effects over centres or regions.

It is common practice to analyse the main and interaction effects in a two-factorial design by applying the two-way analysis of variance (ANOVA). Unfortunately, the corresponding F-tests offer only an overall decision concerning the presence of these effects [Hothorn et al., 2008]. Within this thesis a general methodology to formulate the research hypothesis for an in-depth analysis of interactions is presented. Therefore, the approach presented by Gabriel et al. [1973] is adopted. They constructed product-type interaction contrasts as a direct Kronecker product of two one-way contrasts. These interaction contrasts can be conceived as the difference of differences among means. This idea is further extended in this thesis by formulating the interaction hypotheses as ratio of differences among means. The latter approach has the advantage that the interaction effect is interpretable as relative change of the differences among means. Furthermore, this extension of product-type interaction contrasts allows the differentiation between quantitative and qualitative interactions [Kitsche and Hothorn, 2013]. Hauschke and Kieser [2001] illustrated the usage of multiple ratios of treatment means to test for non-inferiority in medical research. This approach is expanded by using multiple ratios of differences among means to assess non-inferiority of these differences among means. This extension applicable for the assessment of consistency

of the treatment effect in multi-regional trials.

Because in such a detailed analysis several hypotheses are tested simultaneously, an adequate multiple comparison procedure has to be used that controls the family-wise type I error rate for the family of hypotheses under consideration. Multiplicity adjusted p-values for the individual hypotheses are provided, such that the significance of the detailed interpretations can be inferred, while controlling the overall probability of a type I error. Compatible simultaneous confidence intervals can be used to interpret the direction, magnitude and the biological relevance of the interaction effects.

In the first instance the proposed methodologies are presented for factorial experiments where the response variable is assumed to be normally distributed with homogeneous error terms. Furthermore, the situation in which the assumption of homogeneous variances is not fulfilled is considered. In pharmaceutical and biomedical research the outcome of interest is often an "event" with the data taking a binary form commonly denoted as success or failure. Therefore, the developed approach is expanded to the binary case and the hypotheses are formulated as differences of risk differences and ratios of risk differences, respectively.

This thesis is organized as follows. In Chapter 2 five motivating example data sets are presented. The statistical model for normally distributed outcome variables is given in Chapter 3. This chapter also considers the formulation of the interaction effects as the difference of treatment differences and the ratio of treatment differences, respectively. Methods to construct multiple contrast tests and simultaneous confidence intervals to evaluate the different kinds of interactions are illustrated in Chapter 4. Chapter 5 demonstrates the usage of the ratios of treatment differences for the assessment of qualitative interactions. In addition to the proposed method, three commonly used tests to detect qualitative interactions are presented in Chapter 5. In Chapter 6 the research hypothesis of detecting statistical interaction is reversed in order to assess consistency of treatment effects over the levels of a secondary factor. Afterwards, the presented principles are extended for the analysis of interactions in cases of binary response variables in Chapter 7. Chapter 8 presents the properties of the proposed method though a simulation study. In Chapter 9 the motivating examples are analysed. Finally, Chapter 10 provides a concluding discussion.

Chapter 2.

Data Examples

In this chapter five data examples are described where the interest is at least partially on the detection of a statistical interaction. The analysis of the presented example data sets by using the methodologies described within this thesis are given in Chapter 9.

2.1. Bush beans data set

The first example was published by Petersen [1985, p. 155]. The goal of the experiment was to investigate the effect of row spacing on the yield of different varieties of bush beans. Due to the different growth habits of the considered varieties it was assumed that the spacing effect differs between the varieties. The selected four varieties differ such that "New Era" and "Big Green" form low, bushy plants and the two varieties "Little Gem" and "Red Lake" form erect plants with few branches. The chosen row spacing were of 20, 40 and 60 cm between rows. A randomized complete block design with four blocks and 12 plots per block was used. The yield of dried beans in kilograms per plot was determined after harvest time. Figure 2.1 displays the mean yield for each variety-by-spacing combination. It is obvious that the mean yield increases for the varieties "New Era" and "Big Green", which form little, bushy plants, with increasing row spacing. On the other hand, the mean yield decreases for the two varieties "Little Gem" and "Red Lake", which build erect plants, as the spacing increases. The results of the corresponding ANOVA reveal that this interaction between variety and spacing is highly significant (Table 2.1). The significant overall interaction may now be further analysed: What is the

Table 2.1.: ANOVA table and corresponding F-statistics of the bush beans data set from the two-factorial randomized complete block design.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Block	3	341.90	113.97	8.78	< 0.001
Variety	3	1332.56	444.19	34.22	< 0.001
Spacing	2	72.67	36.33	2.80	0.075
Variety:Spacing	6	871.00	145.17	11.18	< 0.001
Residuals	33	428.35	12.98		

difference in yield increase for different spacing between the bushy and tall groups averages? To what extent do the varieties with similar growth type differ in their reaction to spacing?

2.2. Lettuce data set

The second example originates from an experiment that was conducted to analyse the effects of soil type and phosphate fertilizers on lettuce crops. The primary response variable was dry matter in grams per plot. Three different soil types ($S1$, $S2$ and $S3$) and four different levels of phosphate fertilization (including an untreated control) were investigated in a balanced, completely cross-classified treatment structure, laid out as completely randomized design with four replications per treatment combination. The original data for this example were not available and therefore data that reproduce the same treatment means as reported by Bradu and Gabriel [1974] were generated.

Figure 2.2 illustrates the mean dry matters for each fertilizer-by-soil combination. From Figure 2.2 it is obvious that all fertilizers have an increasing effect on dry matter in comparison to the control group for soil type $S1$ and $S3$, whereas the dry matter is almost not affected by the phosphate fertilizers for soil type $S2$. The corresponding ANOVA (Table 2.2) reveals a highly significant interaction between the factors phosphate fertilizer and soil type. The objective is now to compare the three fertilizer effects, defined as the difference of each active fertilization group to the untreated control group, between the different soil types.

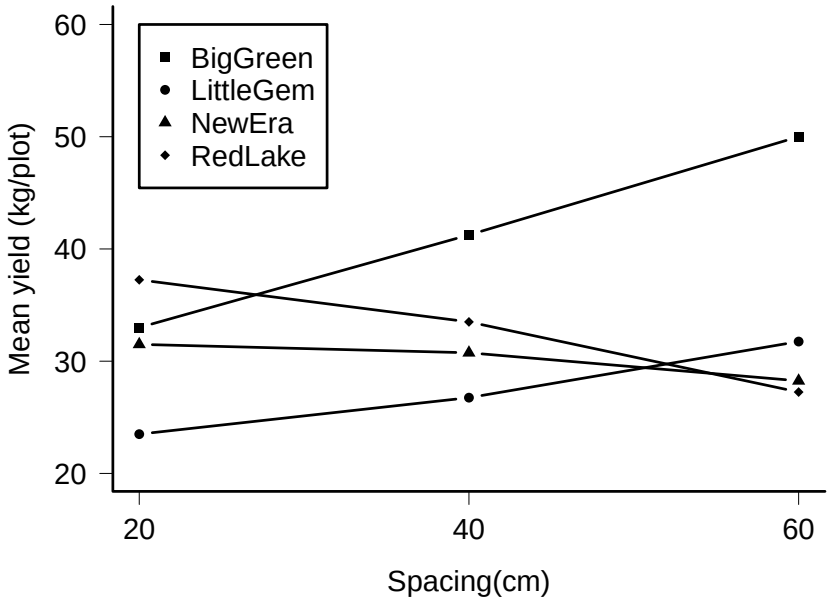

Figure 2.1.: Interaction plot of cell means which illustrates the relationship between row spacing and yield of four bush bean varieties that form either little, bushy plants (New Era and Big Green) or erect plants with few branches (Little Gem and Red Lake).

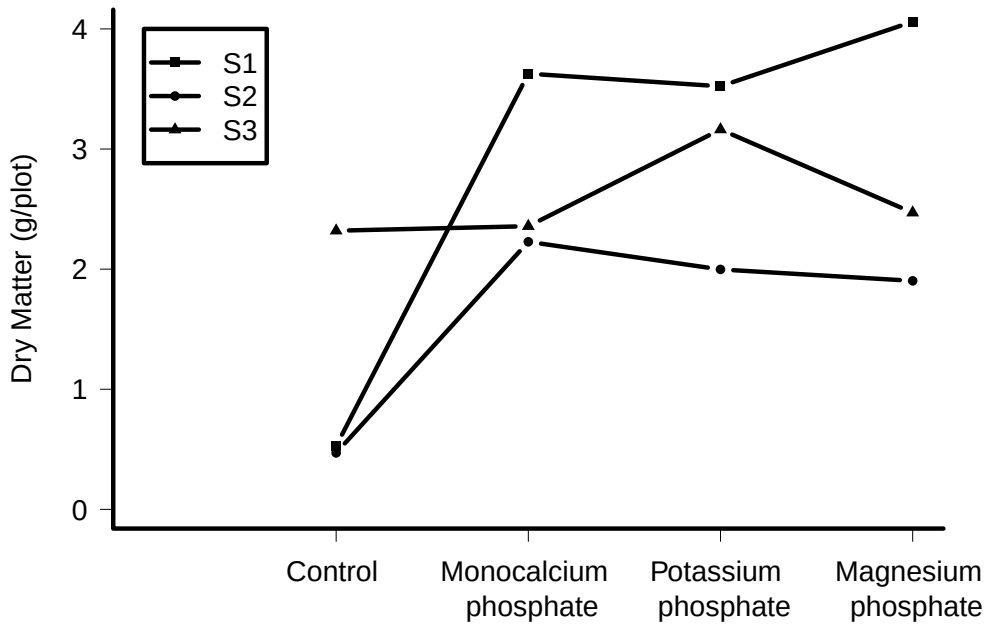

Figure 2.2.: Interaction plot of cell means which illustrates the relationship between phosphate fertilizer and dry matter on lettuce crops for three soil types.

Table 2.2.: ANOVA table and corresponding F-statistics of the lettuce data set from a two-factorial completely randomized design.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fertilizer	3	26.41	8.80	54.99	< 0.001
Soil	2	14.08	7.04	43.98	< 0.001
Fertilizer:Soil	6	14.65	2.44	15.25	< 0.001
Residuals	36	5.76	0.16		

2.3. Multi-centre clinical trial

In this section, an example published in Dmitrienko et al. [2005] is presented. In the multi-centre depression trial, two groups of patients, one treatment and one placebo group, were compared. The primary endpoint was the change from the baseline to the end of the nine week acute treatment phase in the 17-item Hamilton depression rating scale total score (HAMD17 score). The change of scores range from -2 to 28, and therefore, it is assume that this endpoint is approximately normally distributed. The experiment was conducted at five centres. The data are displayed as boxplots in Figure 2.3. From Figure 2.3, it is obvious that there is an increasing treatment effect at centres 100, 102, 103 and 104. However, the treatment effect at centre 101 differs from the remainder of the centres in its sign. The goal is now to decide whether centre 101 represents a qualitative interaction or whether this variation occurs by chance. In Table 2.3 the ANOVA F-statistics is listed. From Table 2.3, it becomes apparent that the global treatment-by-centre interaction ANOVA F-test is significant at $\alpha = 0.05$. The aim is now to determine the source and the type of the significant interaction.

2.4. MERIT-HF study

The fourth example describes a multi-regional clinical trial, namely the Metoprolol Controlled-Release Randomized Intervention Trial in Heart Failure (MERIT-HF) [MERIT-HF Study Group, 1999]. The large scale randomized, double blind, placebo controlled trial was conducted to investigate the treatment effect of adding

Table 2.3.: ANOVA table and corresponding F-statistics for the multi-centre clinical trial.

Source	Df	Sum of Squares	Mean Square	F value	Pr > F
Group	1	888.04	888.04	40.07	<.0001
Centre	4	87.14	21.78	0.98	0.4209
Group:Centre	4	507.45	126.86	5.72	0.0004
Residuals	90	1994.38	22.16		

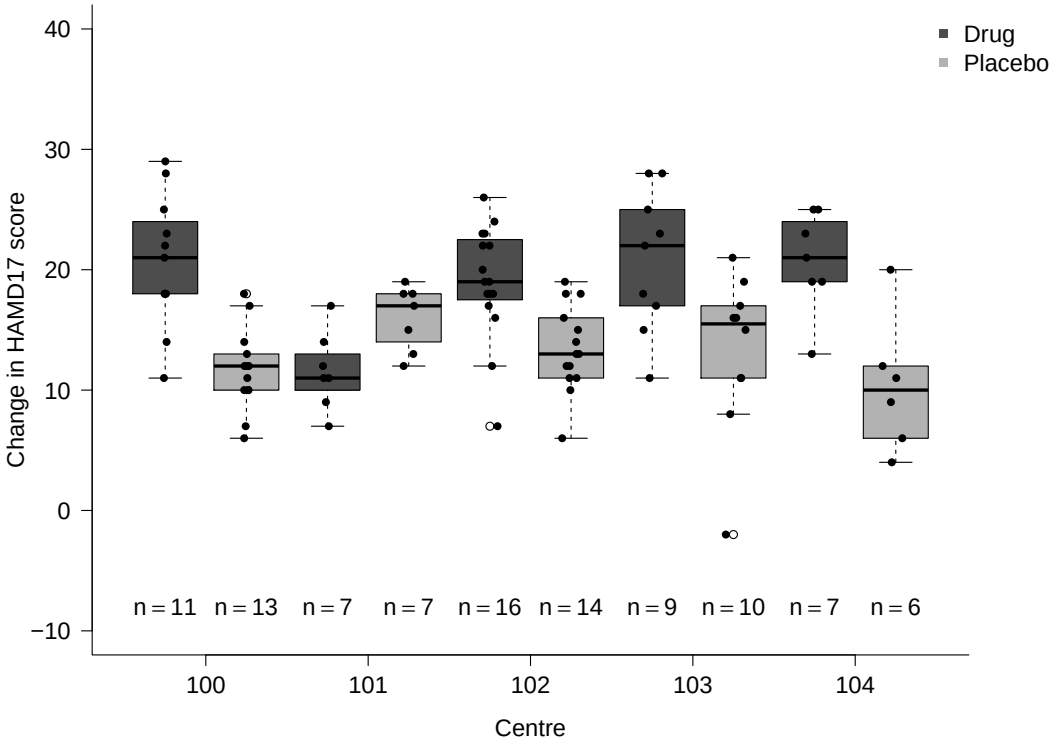

Figure 2.3.: Boxplots of the multi-centre depression trial for each group-by-centre combination. Boxes filled with dark grey represent the experimental drug group, and boxes filled with light grey represent the placebo group. The number of patients for each group-by-centre combination is written below each box. Filled dots denote the data points.

once-daily doses of metoprolol controlled-release/extended-release (Meto CR/XL) to the optimum standard therapy in terms of lowering mortality in patients with symptomatic heart failure. A total number of 3991 patients were randomized into the placebo or the Meto CR/XL group in 14 countries (see Table 2.4). According to Quan et al. [2012], the data from Finland were combined with the data from Denmark, and the data from the Netherlands were combined with the data from Switzerland because no event was observed in the Meto CR/XL group in Finland and Switzerland. From Table 2.4, a decreasing overall treatment effect is observable, whereas in two regions, Iceland and USA, the treatment effect increases. The goal is now to decide if the regions Iceland and USA represent a significant interaction, or if this heterogeneity of the treatment effect occurs by chance only. According to Wedel et al. [2001], significant qualitative interactions were of particular interest, especially significant departures from the overall effect among any of the participating countries. Therefore, in Section 9.4 the focus is on the detection of qualitative interactions.

2.5. Trastuzumab data set

This example presents the data from an interim analysis from a multi-regional clinical trial considering treatment of trastuzumab (Herceptin®, Roche) after adjuvant chemotherapy in HER2-positive breast cancer [Romond et al., 2005]. The international, randomized clinical trial compared two treatment groups, one or two years of trastuzumab being given every three weeks, with observation in patients with HER2-positive breast cancer. The primary endpoint was disease-free survival, defined as time from randomization to the first occurrence of an event, where an event is given as recurrence of breast cancer, contralateral breast cancer, second non-breast malignant disease, or death. The endpoint under consideration is the occurrence of an event because the disease-free survival times were not available. Five regions were pre-specified, namely Central and South America, Eastern Europe, Asia Pacific including Japan and Others (Others include Western and Northern Europe, Canada, South Africa, Australia, New Zealand). In the report of the interim analysis, only the results of the treatment group with trastuzumab treatment for one year and the observation group were presented. The results of

Table 2.4.: Summary table with the number of successes and failures, the total number of observations, and the proportions of successes for each region in the MERIT-HF trial.

Region (j)	Treatment	Outcome		Total	Proportion
		Success	Failure	(n_{ij})	Success
Belgium	Meto CR/XL	3	65	68	0.04
	Placebo	13	53	66	0.20
Czech Republic	Meto CR/XL	9	114	123	0.07
	Placebo	17	107	124	0.14
Denmark/Finland	Meto CR/XL	11	150	161	0.07
	Placebo	13	151	164	0.08
Germany	Meto CR/XL	19	233	252	0.08
	Placebo	31	216	247	0.13
Hungary	Meto CR/XL	16	195	211	0.08
	Placebo	29	183	212	0.14
Iceland	Meto CR/XL	2	17	19	0.11
	Placebo	2	20	22	0.09
Norway	Meto CR/XL	6	91	97	0.06
	Placebo	11	94	105	0.10
Poland	Meto CR/XL	8	94	102	0.08
	Placebo	8	94	102	0.08
Sweden	Meto CR/XL	2	37	39	0.05
	Placebo	9	37	46	0.20
The Netherland/Switzerlnd	Meto CR/XL	14	285	299	0.05
	Placebo	26	265	291	0.09
UK	Meto CR/XL	4	83	87	0.05
	Placebo	9	74	83	0.11
USA	Meto CR/XL	51	481	532	0.10
	Placebo	49	490	539	0.09
Total	Meto CR/XL	145	1845	1990	0.07
	Placebo	217	1784	2001	0.12

Table 2.5.: Summary table with the number of successes and failures, the total number of observations, and the proportions of successes for each region in the Trastuzumab data set.

Region (j)	Treatment	Outcome		Total	Proportion
		Success	Failure	(n_{ij})	Success
Japan only	Trastuzumab	3	38	41	0.073
	Observation	6	40	46	0.130
Asia Pacific, Japan	Trastuzumab	12	190	202	0.059
	Observation	27	175	202	0.134
Eastern Europe	Trastuzumab	10	179	189	0.053
	Observation	26	149	175	0.149
Central and South America	Trastuzumab	7	87	94	0.074
	Observation	8	86	94	0.085
Others	Trastuzumab	98	1110	1208	0.088
	Observation	158	1064	1222	0.129
All regions	Trastuzumab	127	1474	1693	0.075
	Observation	219	1566	1693	0.129

this interim analysis were already discussed in the context of consistency assessment with special focus on the Japanese subgroup by Ando and Hamasaki [2010]. Table 2.5 lists the number of events for each treatment-by-region combination. In addition to the pre-specified regions the data for the Japanese subgroup are given in Table 2.5. The goal is to assess the consistency of the treatment effect, defined as the difference between the trastuzumab treatment group and the observation group, over the participating regions.

Chapter 3.

Hypotheses for statistical interactions

This chapter presents the statistical model under consideration. In addition, a general procedure to formulate appropriate hypotheses to test for statistical interactions via contrasts among means is proposed. Finally, the hypotheses to test for interactions are reformulated as ratio of linear combinations among means.

3.1. The model

For the sake of convenience and without loss of generality, a completely randomized design with two factors, afterwards denoted as A and B is assumed. Nevertheless, the presented approach can be extended to designs with more than two factors. The endpoint of interest is a continuous and normally distributed outcome measure. Furthermore, let I be the number of levels of factor A (with index $i = 1, \ldots, I$), and J be the number of levels of factor B (with index $j = 1, \ldots, J$). The number of experimental units is permitted to vary between the factor combinations and is denoted by n_{ij}. The total number of experimental units is given by $N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. The corresponding two-way ANOVA model with an interaction term is given by:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{3.1}$$

where the parameter μ denotes the grand mean, α_i is the treatment effect for the ith level of factor A, β_j is the treatment effect for the jth level of factor B and $(\alpha\beta)_{ij}$ denotes the joint effect of the ith level of factor A and the jth level of factor B. Furthermore, it is assumed that the error associated with the kth observation, with $k = 1, \ldots, n_{ij}$, is normally distributed with common variance, $\epsilon_{ijk} \sim N(0, \sigma^2)$.

For the purpose of the presented method the ANOVA model is reformulated into the cell means model as follows:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \tag{3.2}$$

where the parameter μ_{ij} denotes the cell mean of the ith level of factor A and the jth level of factor B. Please note, that the parameters of the two models (3.1) and (3.2) can easily be transformed into each other in the following way:

- the main effect $\alpha_i = \mu_{i.} - \mu_{..}$
- the main effect $\beta_j = \mu_{.j} - \mu_{..}$
- the interaction effect $(\alpha\beta)_{ij} = \mu_{ij} - (\mu_{i.} - \mu_{..}) - (\mu_{.j} - \mu_{..}) - \mu_{..} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$,

where the dot notation denotes the mean over the corresponding factor levels. It is well known, that the interaction effect is equal to zero if each cell mean is represented by the additivity of the main effects [Scheffe, 1999]. The vector of cell means is given by the column vector $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_{11}, \ldots, \mu_{I1}, \ldots, \mu_{1J}, \ldots, \mu_{IJ})$, where the elements of $\boldsymbol{\mu}$ are primarily ordered according to factor B, and within factor B according to factor A (the superscripted T on a vector or matrix denotes the transpose). For the sake of clarity the vector $\boldsymbol{\mu}$ is given the new index l, resulting in $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_1, \ldots, \mu_L)$, where $L = I \cdot J$, the number of factor combinations. For later use, the $L \times L$ matrix $\boldsymbol{M}$ defines the diagonal matrix that contains the reciprocals of the sample sizes $\boldsymbol{M} = \mathrm{diag}\left(n_1^{-1}, \ldots, n_L^{-1}\right)$. Furthermore, the maximum likelihood estimator of $\boldsymbol{\mu}$ is denoted by $\hat{\boldsymbol{\mu}}$, with $\hat{\mu}_l = \bar{Y}_l = \sum_{k=1}^{n_l} Y_{lk}/n_l$. The pooled sample variance is given by $s^2 = \sum_{l=1}^{L} \sum_{k=1}^{n_l} \left(Y_{lk} - \bar{Y}_l\right)^2 / (N - L)$. The square root of the pooled variance estimator is denoted by s, i.e, the pooled sample standard deviation.

Table 3.1.: Table of cell means for a two-factorial designs, where factor A has $i = 1, 2, 3, \ldots, I$ levels and factor B has $j = 1, 2, 3, \ldots, J$ levels. Grey marked cells of the $I \times J$ table correspond to the required parameters to calculate the interaction residuals resulting from the direct products of deviations from the grand mean.

	B_1	B_2	B_3	B_j	B_J	Means
A_1	μ_{11}	μ_{12}	μ_{13}	$\cdots$	μ_{1J}	$\mu_{1.}$
A_2	μ_{21}	μ_{22}	μ_{23}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_3	μ_{31}	μ_{32}	μ_{33}	$\cdots$	μ_{3J}	$\mu_{3.}$
A_i	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	μ_{I1}	μ_{I2}	μ_{I3}	$\cdots$	μ_{IJ}	$\mu_{I.}$
Means	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\cdots$	$\mu_{.J}$	$\mu_{..}$

The cell means from a two-way layout can be illustrated in a $I \times J$ table, see e.g., Table 3.1. Each cell in Figure 3.1 corresponds to a cell mean parameter from model (3.2). The lower and right margins in Table 3.1 represent the marginal means averaged over the levels of the other factor.

3.2. Contrasts among means

It is assumed that the research question can be translated into the $m = 1, \ldots, M$ general linear hypotheses according to Searle [1997]:

$$H_0 : \quad \boldsymbol{\psi} := \boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{\theta}, \tag{3.3}$$

where $\boldsymbol{\psi}$ corresponds to the vector of parameters of interest, $\boldsymbol{C}$ defines an $M \times L$ contrast matrix and $\boldsymbol{\theta}$ denotes a vector of dimension L of specified constants. Each row vector $\boldsymbol{c}_m$ in $\boldsymbol{C}$ defines one contrast and corresponds to a single research hypothesis:

$$H_0^{(m)} : \quad \boldsymbol{\psi}^m := \boldsymbol{c}_m\boldsymbol{\mu} = \theta.$$

According to Kirk [1995], a contrast is defined as a linear combination of means with known weights or constants, denoted by c_{ml}. Each element of $\boldsymbol{c}_m$ corresponds to a contrast coefficient c_{ml} associated with the lth parameter in $\boldsymbol{\mu}$ from the mth hypothesis of interest. For the choice of the contrast coefficients the following

restrictions are given: (i) at least one contrast coefficient is not equal to zero ($c_{ml} \neq 0$ for some l), (ii) and the sum of the coefficients in a given contrast is equal to zero ($\sum_{l=1}^{L} c_{ml} = 0$) [Kirk, 1995]. An additional restriction is given, if the researcher wants to compare the magnitude of different contrasts. In this case the values of positive contrast coefficients have to sum to 1 and also the absolute values of negative contrast coefficients have to sum to 1 ($\sum_{l:c_{ml}\leq 0} |c_{ml}| = \sum_{l:c_{ml}\geq 0} c_{ml} = 1$).

To investigate contrasts that correspond to the analysis of interactions, the different types of contrasts which seem to be of special interest for the main effect according to Gabriel et al. [1973] are recalled. The four different types are demonstrated using factor A:

1. deviations from the grand mean: $\mu_{i.} - \mu_{..}$
2. pairwise differences: $\mu_{i.} - \mu_{i'.}$, where $(i \neq i')$
3. dichotomy contrasts: - $\mu_{R.} - \mu_{\bar{R}.}$, where R is any non-empty subset of the set of indices $1, \ldots, I$, and $\bar{R}$ is the complementary subset of size $I - R$.
4. pooled-mean differences: $\mu_{R'.} - \mu_{R''.}$, where R' and R'' are two non-empty disjoint subsets of $1, \ldots, I$.

Several popular contrasts correspond to the classification of Gabriel et al. [1973]. The all-pair comparison of Tukey [Braun, 1994] and the many-to-one comparison of Dunnett [1955] are well known examples for pairwise differences. The trend test of Williams [1971] and its expression as contrast test according to Bretz [2006] are examples for the pooled-mean differences.

3.3. Product-type interaction contrasts

In this section the approach of Gabriel et al. [1973] is adopted, who formulate product-type interaction contrasts to detect interactions in a two-way layout. According to Gabriel et al. [1973], the listed contrasts for the main effects in Section 3.2 are appropriate to define meaningful interaction contrasts. They developed a product-type interaction contrast as a direct Kronecker product, denoted by $\otimes$, of the two "one-way contrasts". Recall that the Kronecker product of two matrices

multiplies each element of the first matrix with the second matrix. Each of the following product-type interaction contrasts is a direct product of the one-way contrasts belonging to the set above. This procedure yields to the following four sets of interaction contrasts:

1. interaction residuals: $\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$ - are direct products of deviations from the grand mean. These interaction effects are obtained by removing row, column and grand mean effects from the cell means. This kind of product-type interaction contrast is also known as "corrected cell means" [Boik, 1993] or "interaction score" [Abelson and Prentice, 1997]. For illustrative purposes, Table 3.1 marks the required cells in the $I \times J$ table of cell means to calculate an interaction residual. The interaction residuals correspond to the interaction effects $(\alpha\beta)_{ij}$ from the two-way ANOVA model in Equation 3.1. This set is of particular interest in situations where it is suspected that the additive model may hold for all but one (or all but a very few) of the cell means μ_{ij}.

2. tetrad contrasts: $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'}$, where $(i \neq i', j \neq j')$, are direct products of pairwise differences. By way of illustration, Table 3.2 considers the first and third level of factor A and the first and third level of factor B. The corresponding local hypothesis to test the interaction effect compares the difference between B_1 and B_3 at A_1 to the difference between B_1 and B_3 at A_3:$(\mu_{11} - \mu_{13}) - (\mu_{31} - \mu_{33}) = \mu_{11} - \mu_{13} - \mu_{31} + \mu_{33}$. This set is also of interest in situations where the additive model is suspected to hold except for a minority of cell means.

3. double-dichotomy contrasts: $\mu_{R\times S} - \mu_{R\times \bar{S}} - \mu_{\bar{R}\times S} + \mu_{\bar{R}\times \bar{S}}$, where R and $\bar{R}$ are complementary subsets of $1, \ldots, I$, and S and $\bar{S}$ are complementary subsets of $1, \ldots, J$; $\mu_{R\times S}$ is the mean of all numbers μ_{ij} with $i \in R$ and $j \in S$; $\mu_{\bar{R}\times \bar{S}}$ is the mean of all numbers μ_{ij} with $i \in \bar{R}$ and $j \in \bar{S}$, etc. Double-dichotomy contrasts are direct products of dichotomy contrasts. This set is of interest in situations where it is suspect that additivity may hold within some submatrices of the $I \times J$ matrix of cell means, but not between the sub-matrix means.

4. pooled-tetrad contrasts: $\mu_{R' \times S'} - \mu_{R' \times S''} - \mu_{R'' \times S'} + \mu_{R'' \times S''}$ - are direct products of pooled mean differences (defined analogously to the double-dichotomy contrasts, except that the disjoint subsets R' and R'', S' and S'', respectively, do not need to be complementary).

To illustrate the construction of appropriate contrasts, the next subsections demonstrate the choice of contrast coefficients to formulate the hypotheses of interest for the examples presented in Section 2.1, 2.2 and 2.3.

Table 3.2.: Schematic display of tetrad contrasts: grey marked cells of the $I \times J$ table correspond to the required parameters to calculate the tetrad contrast of cell means: $\mu_{11} - \mu_{13} - \mu_{31} + \mu_{33}$.

	B_1	B_2	B_3	B_j	B_J	Means
A_1	μ_{11}	μ_{12}	μ_{13}	$\cdots$	μ_{1J}	$\mu_{1.}$
A_2	μ_{21}	μ_{22}	μ_{23}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_3	μ_{31}	μ_{32}	μ_{33}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_i	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	μ_{I1}	μ_{I2}	μ_{I3}	$\cdots$	μ_{IJ}	$\mu_{I.}$
Means	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\cdots$	$\mu_{.J}$	$\mu_{..}$

Table 3.3.: Schematic display of double-dichotomy contrasts: grey marked cells of the $I \times J$ table correspond to the required parameters to calculate the double-dichotomy contrasts resulting from direct products of dichotomy contrasts. The different grey scale defines the complementary subsets S and $\bar{S}$, and R and $\bar{R}$.

	B_1	B_2	B_3	B_j	B_J	Means
A_1	μ_{11}	μ_{12}	μ_{13}	$\cdots$	μ_{1J}	$\mu_{1.}$
A_2	μ_{21}	μ_{22}	μ_{23}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_3	μ_{31}	μ_{32}	μ_{33}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_i	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	μ_{I1}	μ_{I2}	μ_{I3}	$\cdots$	μ_{IJ}	$\mu_{I.}$
Means	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\cdots$	$\mu_{.J}$	$\mu_{..}$

Table 3.4.: Schematic display of pooled-tetrad contrasts: grey marked cells of the $I \times J$ table correspond to the required parameters to calculate the pooled-tetrad contrasts resulting from direct products of pooled mean differences.

	B_1	B_2	B_3	B_j	B_J	Means
A_1	μ_{11}	μ_{12}	μ_{13}	$\cdots$	μ_{1J}	$\mu_{1.}$
A_2	μ_{21}	μ_{22}	μ_{23}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_3	μ_{31}	μ_{32}	μ_{33}	$\cdots$	μ_{2J}	$\mu_{2.}$
A_i	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	μ_{I1}	μ_{I2}	μ_{I3}	$\cdots$	μ_{IJ}	$\mu_{I.}$
Means	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	$\cdots$	$\mu_{.J}$	$\mu_{..}$

3.3.1. Contrasts for the bush beans example

At first, the contrast matrices for the main effects for the factors spacing $\boldsymbol{C}_{\text{Spacing}}$ and variety $\boldsymbol{C}_{\text{Variety}}$ are specified to construct the product-type interaction contrast matrix $\boldsymbol{C}_{\text{Interaction}}$. For the factor spacing the Tukey-type contrasts of all pairwise comparisons between the levels are used (see Equation 3.4). For the factor variety the goal is to first compare the average of the two tall varieties with those of the two bushy varieties (see first row in $\boldsymbol{C}_{\text{Variety}}$ in Equation 3.4). Furthermore, the varieties within each of the two growth types are compared, see row 2 and 3 from $\boldsymbol{C}_{\text{Variety}}$.

$$\boldsymbol{C}_{\text{Variety}} = \begin{pmatrix} 0.5 & 0.5 & -0.5 & -0.5 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \qquad \boldsymbol{C}_{\text{Spacing}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \tag{3.4}$$

The direct Kronecker product of these matrices which results in the matrix $\boldsymbol{C}_{\text{Variety}} \otimes \boldsymbol{C}_{\text{Spacing}} = \boldsymbol{C}_{\text{Interaction}}$ is built to get the following product-type interaction contrasts:

$$\boldsymbol{C}_{\text{Interaction}} = \begin{pmatrix} 0.5 & -0.5 & 0 & 0.5 & -0.5 & 0 & -0.5 & 0.5 & 0 & -0.5 & 0.5 & 0 \\ 0.5 & 0 & -0.5 & 0.5 & 0 & -0.5 & -0.5 & 0 & 0.5 & -0.5 & 0 & 0.5 \\ 0 & 0.5 & -0.5 & 0 & 0.5 & -0.5 & 0 & -0.5 & 0.5 & 0 & -0.5 & 0.5 \\ 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}. \tag{3.5}$$

The first row in $\boldsymbol{C}_{\text{Interaction}}$ compares the difference of the tall and bushy varieties between the first and the second row spacing level. Using the classification of Gabriel et al. [1973] the first three rows in $\boldsymbol{C}_{\text{Interaction}}$ are examples of pooled-tetrad contrasts. Row number 4-6 are examples of tetrad contrasts where only the

sub-matrix of the tall varieties is analysed. The same holds for the rows 7-9 where the sub-matrix of the bushy varieties is analysed in more detail.

3.3.2. Contrasts for the lettuce data set

The factor phosphate fertilizer consists of an untreated control group and three different phosphorus fertilizers. Objective for this factor is to estimate the increase in yield that results applying each of the fertilizers compared to the untreated control. The corresponding Dunnett-type contrast matrix is given by $\boldsymbol{C}_{\text{Phosphate}}$ in Equation 3.6. The second factor, soil type, contains no further sub-structure, and interest is in comparing all three soil types among each other. The contrast matrix $\boldsymbol{C}_{\text{Soil}}$ in Equation 3.6 specifies the Tukey-type contrasts of all pairwise comparisons of the soil types.

$$\boldsymbol{C}_{\text{Phosphate}} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \qquad \boldsymbol{C}_{\text{Soil}} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}. \tag{3.6}$$

The resulting interaction contrast matrix $\boldsymbol{C}_{\text{Interaction}}$ in Equation 3.7 is given by the direct Kronecker product $\boldsymbol{C}_{\text{Phosphate}} \otimes \boldsymbol{C}_{\text{Soil}}$. The contrasts in $\boldsymbol{C}_{\text{Interaction}}$ allow to interpret to what extend the difference in yield between the three phosphate fertilizers compared to the control varies between the three soil types. According to the classification system of Gabriel et al. [1973], the contrasts in $\boldsymbol{C}_{\text{Interaction}}$ are denoted as tetrad contrasts.

$$\boldsymbol{C}_{\text{Interaction}} = \begin{pmatrix} -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \tag{3.7}$$

3.3.3. Contrasts for the multi-centre clinical trial

The contrast matrix $\boldsymbol{C}_{\text{Group}}$ in Equation 3.8 defines the difference between the active treatment group and the placebo group, which is commonly denoted as treatment effect.

When analysing the interaction in a multi-centre clinical trial, the aim is to detect a centre at which the treatment effect differs in accordance with the general treatment effect. Each row in the contrast matrix $\boldsymbol{C}_{\text{Centre}}$ in Equation 3.8 represents a grand mean contrast, which compares one centre (the first centre in the first row, the second centre in the second row, etc.) to the general mean. The contrast matrix $\boldsymbol{C}_{\text{GrandMean}}$ in Equation 3.9 represents the general form of the grand mean contrast matrix.

$$\boldsymbol{C}_{\text{Group}} = \begin{pmatrix} 1 & -1 \end{pmatrix}, \qquad \boldsymbol{C}_{\text{Centre}} = \begin{pmatrix} 0.76 & -0.14 & -0.30 & -0.19 & -0.13 \\ -0.24 & 0.86 & -0.30 & -0.19 & -0.13 \\ -0.24 & -0.14 & 0.70 & -0.19 & -0.13 \\ -0.24 & -0.14 & -0.30 & 0.81 & -0.13 \\ -0.24 & -0.14 & -0.30 & -0.19 & 0.87 \end{pmatrix}. \tag{3.8}$$

$$\boldsymbol{C}_{\text{GrandMean}} = \begin{pmatrix} 1-\frac{n_1}{N} & -\frac{n_2}{N} & \cdots & -\frac{n_J}{N} \\ -\frac{n_1}{N} & 1-\frac{n_2}{N} & \cdots & -\frac{n_J}{N} \\ \vdots & \vdots & 1-\frac{n_j}{N} & \vdots \\ -\frac{n_1}{N} & -\frac{n_2}{N} & \cdots & 1-\frac{n_J}{N} \end{pmatrix} \tag{3.9}$$

The corresponding product-type interaction contrast matrix $\boldsymbol{C}_{\text{Interaction}}$ is given by $\boldsymbol{C}_{\text{Centre}} \otimes \boldsymbol{C}_{\text{Group}}$:

$$\boldsymbol{C}_{\text{Interaction}} = \begin{pmatrix} 0.76 & -0.76 & -0.14 & 0.14 & -0.30 & 0.30 & -0.19 & 0.19 & -0.13 & 0.13 \\ -0.24 & 0.24 & 0.86 & -0.86 & -0.30 & 0.30 & -0.19 & 0.19 & -0.13 & 0.13 \\ -0.24 & 0.24 & -0.14 & 0.14 & 0.70 & -0.70 & -0.19 & 0.19 & -0.13 & 0.13 \\ -0.24 & 0.24 & -0.14 & 0.14 & -0.30 & 0.30 & 0.81 & -0.81 & -0.13 & 0.13 \\ -0.24 & 0.24 & -0.14 & 0.14 & -0.30 & 0.30 & -0.19 & 0.19 & 0.87 & -0.87 \end{pmatrix}. \tag{3.10}$$

An alternative choice for $\boldsymbol{C}_{\text{Centre}}$ would be a contrast matrix in which all of

the pairwise comparisons are made (Tukey-type contrasts). The disadvantage of this choice is that the number of contrasts M rapidly increases with an increasing number of centres ($M = J \cdot (J-1)/2$). This increase results in a confusing decision on the location of the interaction. Therefore, the choice of $\boldsymbol{C}_{\text{Centre}}$ as $\boldsymbol{C}_{\text{GrandMean}}$ is recommended, especially if the number of levels of the secondary factor is high with respect to that of the primary factor, e.g., centres or regions. If the secondary factor has only a few levels, e.g., subgroups, gender or biomarker, then the usage of all pairwise contrasts is recommended.

3.4. Ratios among contrasts of means

In this section the research hypotheses for the analysis of interactions are formulated in terms of ratios of linear combinations of means. This formulation allows the interpretation of the interaction effects as relative changes, e.g. as percentage changes, between differences of means. Djira [2005] presented the formulation of the research hypothesis using several ratios of linear combinations of the treatment means for the one-way ANOVA model with homogeneous variances. Within here, this approach is expanded through the two-way layout using the cell means model from Equation 3.2. In the following, the ratios of linear combinations of means are denoted as $\boldsymbol{\gamma}$ and the global null hypothesis is given by:

$$H_0: \quad \boldsymbol{\gamma} := \frac{\boldsymbol{H}\boldsymbol{\mu}}{\boldsymbol{D}\boldsymbol{\mu}} = \boldsymbol{\omega}, \tag{3.11}$$

where $\boldsymbol{H}$ and $\boldsymbol{D}$ represent the $M \times L$ numerator and denominator contrast matrices, respectively, and $\boldsymbol{\omega}$ denotes a vector of dimension L of specified relative thresholds. Each research hypothesis can now be formulated by the mth row vector of the numerator and denominator contrast matrix, namely $\boldsymbol{h_m}$ and $\boldsymbol{d_m}$:

$$H_0^{(m)}: \qquad \gamma^m := \frac{\boldsymbol{h}_m\boldsymbol{\mu}}{\boldsymbol{d}_m\boldsymbol{\mu}} = \omega.$$

Again, the numerator and denominator interaction contrast matrices are build by using two "one-way" contrast matrices. Suppose the contrast matrices for factor A and factor B are given by $\boldsymbol{C}_A$ and $\boldsymbol{C}_B$. Furthermore, it is assumed that interest

is in the assessment of an A-by-B interaction, i.e. the varying difference between the levels of factor A between the levels of factor B. To achieve an appropriate numerator and denominator interaction contrast matrix, the researcher has to built a numerator and denominator contrast matrix for the secondary factor B, i.e. $\boldsymbol{C}_B^{\text{Numerator}}$ and $\boldsymbol{C}_B^{\text{Denominator}}$. Building the direct Kronecker product of these matrices with the contrast matrix of the primary treatment factor $\boldsymbol{C}_A$ results in the $M \times L$ numerator and denominator interaction contrast matrices $\boldsymbol{C}_{\text{Interaction}}^{\text{Numerator}} = \boldsymbol{C}_B^{\text{Numerator}} \otimes \boldsymbol{C}_A$ and $\boldsymbol{C}_{\text{Interaction}}^{\text{Denominator}} = \boldsymbol{C}_B^{\text{Denominator}} \otimes \boldsymbol{C}_A$. The application of this procedure for the multi-centre clinical trial example is given in the next subsection.

3.4.1. Numerator and denominator contrast matrices for the multi-centre clinical trial

In this subsection the hypotheses to detect a statistical interaction in the multi-centre clinical trial example presented in Section 2.3 are reformulated as the ratio of treatment differences. Therefore, the contrast matrix of the primary treatment factor $\boldsymbol{C}_{\text{Group}}$ from Equation 3.8 is used. In addition, the numerator and denominator contrast matrices for the secondary factor centre are defined by Equation 3.12.

$$\boldsymbol{C}_{\text{Centre}}^{\text{Numerator}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \boldsymbol{C}_{\text{Centre}}^{\text{Denominator}} = \begin{pmatrix} 0.24 & 0.14 & 0.30 & 0.19 & 0.13 \\ 0.24 & 0.14 & 0.30 & 0.19 & 0.13 \\ 0.24 & 0.14 & 0.30 & 0.19 & 0.13 \\ 0.24 & 0.14 & 0.30 & 0.19 & 0.13 \\ 0.24 & 0.14 & 0.30 & 0.19 & 0.13 \end{pmatrix} \tag{3.12}$$

The ratio of the numerator and denominator contrast matrices for the secondary factor build the relative effect of each centre to the overall centre effect. The resulting numerator and denominator interaction contrast matrices are given in Equation 3.13 and 3.14.

$$\boldsymbol{H} = \boldsymbol{C}_{\text{Centre}}^{\text{Numerator}} \otimes \boldsymbol{C}_{\text{Group}} =$$

$$\boldsymbol{C}_{\text{Interaction}}^{\text{Numerator}} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \tag{3.13}$$

$$\boldsymbol{D} = \boldsymbol{C}_{\text{Centre}}^{\text{Denominator}} \otimes \boldsymbol{C}_{\text{Group}} =$$

$$\boldsymbol{C}_{\text{Interaction}}^{\text{Denominator}} = \begin{pmatrix} 0.24 & -0.24 & 0.14 & -0.14 & 0.30 & -0.30 & 0.19 & -0.19 & 0.13 & -0.13 \\ 0.24 & -0.24 & 0.14 & -0.14 & 0.30 & -0.30 & 0.19 & -0.19 & 0.13 & -0.13 \\ 0.24 & -0.24 & 0.14 & -0.14 & 0.30 & -0.30 & 0.19 & -0.19 & 0.13 & -0.13 \\ 0.24 & -0.24 & 0.14 & -0.14 & 0.30 & -0.30 & 0.19 & -0.19 & 0.13 & -0.13 \\ 0.24 & -0.24 & 0.14 & -0.14 & 0.30 & -0.30 & 0.19 & -0.19 & 0.13 & -0.13 \end{pmatrix} \tag{3.14}$$

The first row vectors of $\boldsymbol{C}_{\text{Interaction}}^{\text{Numerator}}$ and $\boldsymbol{C}_{\text{Interaction}}^{\text{Denominator}}$ build the ratio of the treatment effect of the first centre through the overall treatment effect. Therefore the parameter γ_1 can be interpreted as the relative change of the treatment effect of the first centre to the overall treatment effect.

Chapter 4.

Inference to test statistical interactions

Within this chapter inferential procedures to test for statistical interactions are presented. At first, the global ANOVA F-test is shortly reviewed. Afterwards, appropriate methods to simultaneously test the M hypotheses defined in Chapter 3 are presented in more detail.

4.1. Global test for statistical interaction

Given a two-factorial design as presented in Section 3.1 it is common practice to evaluate the interaction effect via ANOVA techniques, whereas the ANOVA partitions the total sum of squares into component sum of squares. When applying the classical ANOVA to test for interaction, the null and alternative hypotheses are defined as [Kirk, 1995]:

$$\begin{aligned} H_0: &\quad (\alpha\beta)_{ij} = 0 \quad \text{(for all i and j)} \quad \text{or} \quad \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0 \quad \text{(for all i and j)} \\ H_A: &\quad (\alpha\beta)_{ij} \neq 0 \quad \text{(for all i and j)} \quad \text{or} \quad \mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} \neq 0 \quad \text{(for all i and j).} \end{aligned} \tag{4.1}$$

The test statistic is given by

$$F(\alpha\beta) = \frac{MS(\alpha\beta)}{MS(Error)} \tag{4.2}$$

, where $MS(\alpha\beta)$ denotes the mean square of the interaction effect and $MS(Error)$ the mean square of the error. For a balanced design $MS(\alpha\beta)$ and $MS(Error)$ are given by:

$$MS(\alpha\beta) = \frac{N \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2}{(I-1)(J-1)},$$
$$MS(Error) = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - Y_{ij.})^2}{IJ(N-1)}.$$

The test statistic $F(\alpha\beta)$ follows a F-distribution with $(I-1)(J-1)$ numerator degrees of freedom and $IJ(N-1)$ denominator degrees of freedom [Scheffe, 1999].

It is well known, that the ANOVA F-test only provides global inference on the presence of any effect [Hothorn et al., 2008]. Furthermore, due to the quadratic form of the ANOVA F-statistic is is not possible to make one-sided test decisions [Konietschke et al., 2013]. Since the objective of this thesis is an in depth analysis of statistical interaction the focus is now on the simultaneous assessment of the pre-defined interaction hypotheses presented in Chapter 3.

4.2. Multiple contrast tests for product-type interaction contrasts

The objective is now to simultaneously test the M hypotheses represented by the M rows of a given matrix of interaction contrasts $\boldsymbol{C}_{AB}$. Recall that the global null hypothesis in Equation 3.3 is the intersection of the local null hypotheses $H_0^{(m)}$:

$$H_0 : (\psi_1 = \theta \cap \psi_2 = \theta \cap \ldots \cap \psi_m = \theta) \tag{4.3}$$

and

$$H_A : (\psi_1 \neq \theta \cup \psi_2 \neq \theta \cup \ldots \cup \psi_m \neq \theta). \tag{4.4}$$

To test the global null hypothesis in Equation (4.3), a Union-Intersection test is required, where the global null hypothesis is rejected if any local null hypothesis $H_0^m : \psi_m = \theta$ is rejected.

It is well known that if the M multiple hypotheses are tested simultaneously, the type I error rate increases beyond the a priori defined significance level α. Therefore a multiple comparison procedure is needed that controls a multiple type I error rate to make valid conclusions over the M hypotheses under consideration. One version of the type I error rate in simultaneous testing several hypotheses is the definition of the family-wise error rate (FWER). According to Dmitrienko and D'Agostino [2013] the FWER is controlled in the strong sense at significance level α if "*the probability of erroneously rejecting at least one true null hypothesis is not greater than α, for all possible configurations of true and false null hypotheses*". Within this thesis the concept of multiple contrast tests (MCTs) is used as multiple comparison procedure [Mukerjee et al., 1987, Bretz, 1999]. This allows the simultaneous testing of the M hypotheses defined by $\boldsymbol{C\mu} = \boldsymbol{\theta}$ and to control the FWER in the strong sense.

The test statistic for the mth contrast of cell means is given by the standardized estimator [Bretz, 1999]:

$$T_m = \frac{\sum_{l=1}^{L} c_l \hat{\mu}_l - \theta_m}{s\sqrt{\sum_{l=1}^{L} c_l^2/n_l}} = \frac{\boldsymbol{c}_m \hat{\boldsymbol{\mu}} - \theta_m}{s\sqrt{\boldsymbol{c}_m \boldsymbol{M} \boldsymbol{c}_m^T}}. \tag{4.5}$$

To simultaneous test the M hypotheses the M-dimensional vector of test statistics $\boldsymbol{T} = (T_1, \ldots, T_M)^T$ is considered. Under the null hypothesis the vector of test statistics $\boldsymbol{T}$ follows a central M-variate t-distribution $Mt_{df=\nu,\boldsymbol{R}}$ with ν degrees of freedom and a correlation matrix $\boldsymbol{R}$, see e.g., Genz and Bretz [1999] and Hothorn et al. [2008]. Computational methods for the multivariate t-distribution were presented by Genz and Bretz [2009] and are available in the add-on package `mvtnorm` [Genz et al., 2013] of the statistical software R [R Core Team, 2013]. The degrees of freedom are given by $\nu = \sum_{l=1}^{L}(n_l - 1)$. Each element of the correlation matrix $\boldsymbol{R} = [\rho_{mm'}]$ can be described by

$$\rho_{mm'} = \frac{\sum_{l=1}^{L} c_{ml} c_{m'l}/n_l}{\sqrt{\left(\sum_{l=1}^{L} c_{ml}^2/n_l\right)\left(\sum_{l=1}^{L} c_{m'l}^2/n_l\right)}} \tag{4.6}$$

(for a detailed proof see [Bretz, 1999, page 25]). The null hypothesis for a particular interaction contrast m is rejected if the corresponding absolute value

of the test statistic is greater or equal than a critical point: $|T_m| \geq q_{1-\alpha/2,\nu,R}$, where $q_{1-\alpha/2,\nu,R}$ is the 1-$\alpha/2$ equi-coordinate percentage point from a multivariate t-distribution $Mt_{df=\nu,\boldsymbol{R}}$. The associated p-values can be calculated as $p_m = P(t \geq |T_m|)$, where the variable t follows a multivariate t-distribution $Mt_{df=\nu,\boldsymbol{R}}$.

4.3. Simultaneous confidence intervals for product-type interaction contrasts

In this section simultaneous confidence intervals (SCIs) that are compatible to the MCTs presented before are considered. Within here, SCI are considered because they provide some information on the direction, the magnitude, and the biological relevance of the interaction effects additionally to the test decisions. Furthermore, some regulatory guidelines recommend the usage of SCI instead of multiplicity adjusted p-vales, e.g., the ICH E9 guideline states: "*Estimates of treatment effects should be accompanied by confidence intervals, whenever possible, ...* " [ICH E9, 1998].

Compatible 1-α SCIs for the parameters of interest $\boldsymbol{\psi}$ are given by:

$$\sum_{l=1}^{L} c_{ml}\hat{\mu}_l \pm q_{1-\alpha/2,M,R} \cdot s\sqrt{\sum_{l=1}^{L} \frac{c_{ml}^2}{n_l}}. \tag{4.7}$$

To get compatible test decision one rejects the local null hypothesis H_0^m if the confidence limits do not include θ_m. In addition, the distance of the confidence limit from θ_m is interpretable as a shift on the scale of the response variable.

4.3.1. Heterogeneous variances

In Section 3.1 it was assumed, that the response variable is normally distributed with a common error term over the groups, $\epsilon_{ijk} \sim N(0, \sigma^2)$. Nevertheless, the problem of heteroscedasticity is not uncommon. Therefore, the notation is extended and it is now assumed that the observations are independently normal with mean μ_l and variance σ_l^2, $Y_{lk} \sim N(\mu_l, \sigma_l^2)$. The sample variances are given by $s_l^2 = \sum_{k=1}^{n_l} \left(Y_{lk} - \bar{Y}_l\right)^2 / (n_l - 1)$.

Hasler [2009] presented an approach to construct SCIs and to perform MCTs in the presence of heteroscedasticity in the one-way layout. Within here this method is adopted to test for product-type interactions in the two-way layout. The test statistic for the mth interaction contrast is now given by

$$T_m^* = \frac{\sum_{l=1}^{L} c_l \hat{\mu}_l - \theta_m}{\sqrt{\sum_{l=1}^{L} c_l^2 s_l^2 / n_l}} = \frac{\boldsymbol{c}_m \hat{\boldsymbol{\mu}} - \theta_m}{\sqrt{\boldsymbol{c}_m \hat{\boldsymbol{V}} \boldsymbol{M} \boldsymbol{c}_m^T}}, \tag{4.8}$$

where $\hat{\boldsymbol{V}}$ is a $L \times L$ diagonal matrix that contains the estimated variances $\hat{\boldsymbol{V}} = \text{diag}\,(s_1^2, \ldots, s_L^2)$. Under the null hypothesis each test statistic T_m^* follows a t-distribution with $\hat{\nu}_m$ degrees of freedom, whereas $\hat{\nu}_m$ is given by

$$\hat{\nu}_m = \frac{\left(\sum_{l=1}^{L} \frac{c_{ml}^2 s_l^2}{n_l} \right)^2}{\sum_{l=1}^{L} \frac{c_{ml}^4 s_l^4}{n_l^2 (n_l - 1)}},$$

(for a detailed proof see [Hasler, 2009, page 37]). The corresponding vector of test statistics $\boldsymbol{T}^* = (T_1^*, \ldots, T_M^*)^T$ does not follow a joint m-variate t-distribution in the presence of heteroscedasticity. To overcome this problem, Hasler [2009] proposed a method that uses m distinct m-variate t-distributions. Each test statistic T_m^* is related to a m-variate t-distribution with $\hat{\nu}_m$ degrees of freedom and a correlation matrix $\boldsymbol{R}^* = [\rho_{mm'}^*]$, whose elements are given by

$$\rho_{mm'}^* = \frac{\sum_{l=1}^{L} c_{ml} c_{m'l} s_l^2 / n_l}{\sqrt{\left(\sum_{l=1}^{L} c_{ml}^2 s_l^2 / n_l \right) \left(\sum_{l=1}^{L} c_{m'l}^2 s_l^2 / n_l, \right)}}.$$

Each H_0^m is now rejected if $|T_m^*| \geq q_{1-\alpha/2, \hat{\nu}_m, R^*}$, where $q_{1-\alpha/2, \hat{\nu}_m, R^*}$ is the $1 - \alpha/2$ level equi-coordinate percentage point of a m-variate t-distribution with estimated correlation matrix $\boldsymbol{R}^*$ and $\hat{\nu}_m$ degrees of freedom.

4.4. Multiple contrast tests for ratios of treatment differences

In this section the focus is on simultaneous inference procedures on the ratios among contrasts of means presented in Section 3.4. Within this thesis the methodology proposed by Djira and Hothorn [2009] is used. The central idea from Djira and Hothorn [2009] is based on the reformulation of the ratio problem $\boldsymbol{h}_m\boldsymbol{\mu}/\boldsymbol{d}_m\boldsymbol{\mu} = \omega$ as a linear form $L_m = (\boldsymbol{h}_m - \omega\boldsymbol{d}_m)\boldsymbol{\mu}$, which was first proposed by Fieller [1954]. The estimate of the variance of L_m is given by

$$s^2_{L_m} = s\sqrt{(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)\boldsymbol{M}(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)^T}$$

According to Djira [2005] the test statistic is given by:

$$T_m = \frac{(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)\hat{\boldsymbol{\mu}}}{s\sqrt{(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)\boldsymbol{M}(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)^T}}. \tag{4.9}$$

Under H_0, each local test statistic T_m follows a t-distribution with ν degrees of freedom, whereas the random vector of test statistics $\boldsymbol{T} = (T_1, \ldots, T_M)^T$ jointly follows a central m-variate t-distribution with ν degrees of freedom and a correlation matrix $\boldsymbol{R} = [\rho_{mm'}]$. Each element of $\boldsymbol{R}$ can be described by

$$\rho_{mm'} = \frac{(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)\boldsymbol{M}(\boldsymbol{h}_{m'} - \omega\boldsymbol{d}_{m'})^T}{\sqrt{(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)\boldsymbol{M}(\boldsymbol{h}_m - \omega\boldsymbol{d}_m)^T}\sqrt{(\boldsymbol{h}_{m'} - \omega\boldsymbol{d}_{m'})\boldsymbol{M}(\boldsymbol{h}_{m'} - \omega\boldsymbol{d}_{m'})^T}}, \tag{4.10}$$

where $m \neq m'$ [Djira, 2005]. The global null hypothesis in Equation (3.11) is rejected if $|T_m| \geq q_{1-\alpha/2,\nu,R}$ for at least one m $(m = 1, \ldots, M)$. The critical point $q_{1-\alpha/2,\nu,R}$ is the $1-\alpha/2$ level equi-coordinate percentage point of a m-variate t-distribution with the correlation matrix $\boldsymbol{R}$ and ν degrees of freedom. The associated adjusted p-values can be calculated as $p_m = P(t \geq |T_m|)$, where the variable t follows a multivariate t-distribution $Mt_{df=\nu,\boldsymbol{R}}$ with ν degrees of freedom and the correlation matrix $\boldsymbol{R}$.

4.5. Simultaneous confidence intervals for ratios of treatment differences

As an alternative to adjusted p-values to test the global null hypothesis in Equation (3.11), the goal is to construct SCIs for the ratios of linear combinations of cell means γ_m. Dilba et al. [2006a] presented an approach to determine approximate SCIs for the ratios of linear combinations of the means in the one-way layout based on the multivariate t-distribution. Here, their method is adopted for the problem of determining ratios of linear combinations of treatment means in a two-way layout to detect interactions. The statistic is defined by

$$T_m(\gamma_m) = \frac{(\boldsymbol{h}_m - \gamma_m \boldsymbol{d}_m)\,\boldsymbol{\mu}}{s\sqrt{(\boldsymbol{h}_m - \gamma_m \boldsymbol{d}_m)\,\boldsymbol{M}\,(\boldsymbol{h}_m - \gamma_m \boldsymbol{d}_m)^T}}. \tag{4.11}$$

Jointly, the vector of test statistics $\boldsymbol{T}(\boldsymbol{\gamma}) = (T_1(\gamma_1), \ldots, T_m(\gamma_m))^T$ follows a m-variate t-distribution with ν degrees of freedom and a correlation matrix $\boldsymbol{R}(\gamma_1, \ldots, \gamma_m)$. The correlation coefficients of $\boldsymbol{R}(\gamma_1, \ldots, \gamma_m)$ are similar to Equation (4.10), but ω is replaced by γ_m and $\gamma_{m'}$. It is obvious that the correlation between two contrasts depends now on the known constants $\boldsymbol{h}$ and $\boldsymbol{d}$, the sample sizes n_l and the unknown ratios $\boldsymbol{\gamma}$. Dilba et al. [2004] proposed a plug-in approach, where the unknown ratios in $\boldsymbol{R}(\gamma_1, \ldots, \gamma_m)$ are replaced by its maximum likelihood estimators $\hat{\gamma}_m = \boldsymbol{h}_m\hat{\boldsymbol{\mu}}/\boldsymbol{d}_m\hat{\boldsymbol{\mu}}$. The two-sided Fieller confidence interval for the mth ratio γ_m is the smallest solution of the quadratic equation in γ_m from $T_m^2(\gamma_m) = t_{\alpha/2}^2(\nu)$. Only if the denominator $\boldsymbol{d}_m\hat{\boldsymbol{\mu}}$ is significantly different from 0, the solution of the quadratic equation will lead to a finite value for the confidence limit (for details, see Dilba et al. [2006a] and Subsection 4.5.2).

4.5.1. Heterogeneous variances

In this section the assumption of homogeneous variances is again relaxed and it is assumed that the outcome measures are independently normal with mean μ_l and variance σ_l^2, $Y_{lk} \sim N(\mu_l, \sigma_l^2)$. Hasler [2009] presented an approach to construct SCIs and to perform MCTs for the ratios of means; this method adjusts for heteroscedastic data. The test statistic for the mth ratio of means is now given

by

$$T_m^* = \frac{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\boldsymbol{\mu}}{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)^T}. \tag{4.12}$$

According to Hasler [2009] under H_0^m, the test statistic T_m^* approximately follows a t-distribution with $\hat{\nu}_m$ degrees of freedom, where

$$\hat{\nu}_m = \frac{\left(\sum_{l=1}^{L} \frac{(h_{ml} - \omega d_{ml})^2 s_l^2}{n_l}\right)^2}{\sum_{l=1}^{L} \frac{(h_{ml} - \omega d_{ml})^4 s_l^4}{n_l^2 (n_l - 1)}}. \tag{4.13}$$

The corresponding vector of test statistics $\boldsymbol{T^*} = (T_1^*, \ldots, T_M^*)^T$ does not follow a joint m-variate t-distribution. To overcome this problem, Hasler [2009] proposed a method that uses m distinct m-variate t-distributions. Each test statistic T_m^* is related to a m-variate t-distribution with $\hat{\nu}_m$ degrees of freedom and a correlation matrix $\boldsymbol{R}^* = [\rho_{mm'}^*]$, whose elements are given by

$$\rho_{mm'}^* = \frac{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_{m'} - \omega \boldsymbol{d}_{m'})^T}{\sqrt{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)^T}\sqrt{(\boldsymbol{h}_{m'} - \omega \boldsymbol{d}_{m'})\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_{m'} - \omega \boldsymbol{d}_{m'})^T}}. \tag{4.14}$$

Each H_0^m is then rejected if $|T_m^*| \geq q_{1-\alpha/2,\hat{\nu}_m,R^*}$, where $q_{1-\alpha/2,\hat{\nu}_m,R^*}$ is the $1-\alpha/2$ level percentage point of a m-variate t-distribution with the estimated correlation matrix $\boldsymbol{R}^*$ and $\hat{\nu}_m$ degrees of freedom. Appropriate simultaneous confidence intervals for the heteroscedastic case can be derived as in Section 4.5 using the estimator $\hat{\gamma}_m = \boldsymbol{h}_m\hat{\boldsymbol{\mu}}/\boldsymbol{d}_m\hat{\boldsymbol{\mu}}$ instead of ω in Equation (4.12), (4.13) and (4.14) [Hasler, 2009].

For the calculation of the simultaneous confidence limits and the multiplicity adjusted p-values, the add-on package *mratios* Dilba et al. [2007] from the statistical software R [R Core Team, 2013] can be used.

4.5.2. Fieller type confidence intervals: a geometric representation

In this subsection the geometric representation of the Fieller type confidence intervals [Fieller, 1954] for the ratio parameter γ as defined in Section 3.4 is given. This geometric representation is used to illustrate the scenarios in which the boundaries

of the intervals are not defined. Hirschberg and Lye [2010a] proposed two geometric representations of the Fieller confidence intervals for the ratio of regression parameter estimates. In the following their approach is used for the geometric representation of the intervals of the ratios γ of linear combinations of the parameters from the cell means model as defined in Section 3.4 by:

$$\boldsymbol{\gamma} = \frac{\boldsymbol{H\mu}}{\boldsymbol{D\mu}}.$$

For the purpose of illustration the focus is on one ratio of linear combinations of cell means, $\gamma^m = \frac{\boldsymbol{h}_m\boldsymbol{\mu}}{\boldsymbol{d}_m\boldsymbol{\mu}}$. Figure 4.1 presents the parameter space for the estimates of the numerator $\boldsymbol{h}_m\hat{\boldsymbol{\mu}}$ and denominator $\boldsymbol{d}_m\hat{\boldsymbol{\mu}}$ (this corresponds to the parameter space displayed in Figure 5.1 for two treatment effects δ_1 and δ_2).

According to Hirschberg and Lye [2010b] the ratio and its Fieller type confidence intervals are defined as follows: The ratio γ^m can be displayed as the slope of the line (red line in Figure 4.1) which passes though the origin $(0, 0)$ and the point $(\boldsymbol{h}_m\hat{\boldsymbol{\mu}}, \boldsymbol{d}_m\hat{\boldsymbol{\mu}})$ (red point in Figure 4.1). The value for $\gamma = \frac{\delta_2}{\delta_1}$ is depicted as the slope of this line: the intersection of the line with the vertical line $x = 1$. Next a wedge (grey scaled area in Figure 4.1) is constructed that contains the $100 \cdot (1 - \alpha)\%$ confidence ellipse for a combination of the parameters in γ. The intersection of this wedge with the line $x = 1$ corresponds to the lower and upper confidence limit (for details see Hirschberg and Lye [2010a]).

Figure 4.1 illustrates three (A, B and C) scenarios for the construction of Fieller type confidence intervals. Situation **A** considers the case where the ellipse lies completely on one side of the y-axis. The corresponding wedge does not contain the y-axis resulting in finite confidence limits. As stated in Subsection 4.5 meaningful confidence limits are not defined if the denominator is not significantly different from zero. This corresponds to the null hypothesis $H_0 : \boldsymbol{d}_m\boldsymbol{\mu} = 0$. In those situations the ellipse cuts the y-axis. Plot **B** in Figure 4.1 considers the case in which the ellipse cuts the y-axis but does not contain the origin. The confidence region is again constructed as the intersection of the wedge with the line $x = 1$. The confidence region is bounded but has a small hole in the middle. Obviously, this case occurs if the denominator is close to zero ($H_0 : \boldsymbol{d}_m\boldsymbol{\mu} = 0$ is not rejected), but the numerator is far from zero. Intuitively, the magnitude of the ratio can get

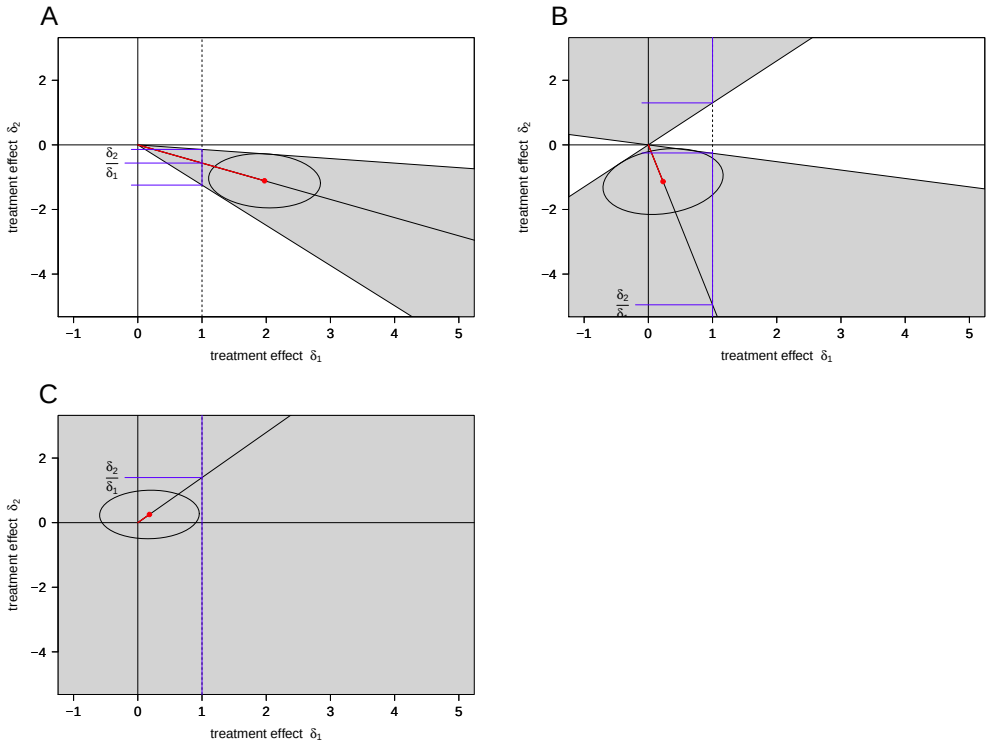

Figure 4.1.: The geometric representation of the Fieller confidence intervals for $\gamma = \delta_2/\delta_1$. The confidence bounds are defined by the rays from the origin that are tangent to the ellipse that defines the 100 $(1-\alpha)$ ellipse for a combination of the parameters in γ. **A**: Case where the lower and upper confidence limits are defined; the confidence ellipse does not intersect the y-axis. **B**: Case where no meaningful limits are defined; the confidence ellipse intersects the y-axis but does not contain the origin. **C**: Case where no Fieller confidence bounds are defined; the confidence ellipse contains the origin.

arbitrary large. Otherwise, there is little confidence about the sign of the ratio because there is little certainty about the sign of the denominator (close to zero). Thus it makes sense that the confidence region is of the form $]-\infty, c_{\text{lower}}]$ and $[c_{\text{upper}}, \infty[$. The small hole in the middle reflects the certainty that a very small numerator or a very large denominator are very unlikely. Situation **C** in Figure 4.1 corresponds to the case where the origin is contained in the ellipse. In this case it is impossible to construct Fieller type confidence intervals since there are no tangents from the origin through the ellipse that can build the wedge. This case occurs if both the numerator and the denominator are not significantly different from zero.

Chapter 5.

Detecting qualitative interactions

As described in Section 1 the detection of qualitative interactions is of particular interest in many research areas [Gail and Simon, 1985]. Within this chapter a method proposed by Kitsche and Hothorn [2013] that uses the ratio of treatment differences is presented to detect a qualitative interaction. Moreover, three global tests to detect a qualitative interaction are illustrated.

5.1. Global tests for qualitative interaction

In this section some frequently applied global tests for qualitative interactions, which were proposed by Azzalini and Cox [1984], Gail and Simon [1985] and Piantadosi and Gail [1993] are presented. Apart from these test several others were published later by Ciminera et al. [1993], Pan and Wolfe [1997] and Li and Chan [2006]. Truberg and Hühn [2000] give a comparative study of different parametric and non-parametric tests to detect qualitative interactions in the context of genotype-by-environment interactions.

5.1.1. Azzalini and Cox test

Azzalini and Cox [1984] presented an approach to test the null hypothesis of no qualitative interaction. The Azzalini and Cox test calculates all pairwise differences of one factor over the levels of the other factor, i.e. all possible tetradic contrasts. The critical value of the test can be obtained from Equation 9 in their paper [Azzalini and Cox, 1984] as

$$t_\alpha = -\Phi^{-1}\left[\left\{-\frac{2\,\log(1-\alpha)}{I(I-1)J(J-1)}\right\}^{\frac{1}{2}}\right],$$

where Φ is the standard normal distribution and α is a pre-specified significance level. Using this critical value a indicator value can be calculated from Equation 5.1.

$$I_{ii'jj'}\begin{cases} = 1 \;\text{ if }\; \mu_{ij} - \mu_{ij'} > \;\; t_\alpha\, \mathrm{s}_{ij}\sqrt{2} \;\text{ and }\; \mu_{i'j} - \mu_{i'j'} < \; -t_\alpha\, \mathrm{s}_{ij}\sqrt{2}, \\ = 1 \;\text{ if }\; \mu_{ij} - \mu_{ij'} < \; -t_\alpha\, \mathrm{s}_{ij}\sqrt{2} \;\text{ and }\; \mu_{i'j} - \mu_{i'j'} > \;\; t_\alpha\, \mathrm{s}_{ij}\sqrt{2}, \\ = 0 \;\text{ otherwise,} \end{cases} \tag{5.1}$$

where s_{ij} is a consistent estimate of the standard error for the mean of each factor combination μ_{ij}. The corresponding decision rule is: reject H_0 of no qualitative interaction if the indicator $I_{ii'jj'} = 1$ for one or more comparisons.

5.1.2. Gail and Simon test

Gail and Simon [1985] proposed a likelihood ratio test for the null hypothesis of no qualitative interaction. They formulated their hypothesis of no qualitative interaction on the basis of the vector of treatment effects $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_J\}$ with $\delta_j = \mu_{1j} - \mu_{2j}$. Therefore, their approach is limited to situations where the primary treatment factor has two levels. These authors demonstrated that this hypothesis implies that the vector of treatment effects $\boldsymbol{\delta}$ lies either in the orthant in which all of the components are non-negative ($\boldsymbol{O}^+ = \{\boldsymbol{\delta} : \delta_j \geq 0 \text{ for all } j\}$) or in the orthant in which all of the components are non-positive ($\boldsymbol{O}^- = \{\boldsymbol{\delta} : \delta_j \leq 0 \text{ for all } j\}$) (see also Figure 5.1):

$$H_0 : \boldsymbol{\delta} \in \boldsymbol{O}^+ \cup \boldsymbol{O}^-. \tag{5.2}$$

The likelihood ratio test proposed by Gail and Simon [1985] to test Equation (5.2) is based on the statistic:

$$\frac{\max_{\boldsymbol{\delta}\in\boldsymbol{O}^+\cup\boldsymbol{O}^-}\exp\sum_{j=1}^{J}\left(-(\hat{\delta}_j - \delta_j)^2/(2s(\delta)_j^2)\right)}{\max_{\boldsymbol{\delta}}\exp\sum_{j=1}^{J}\left(-(\hat{\delta}_j - \delta_j)^2/(2s(\delta)_j^2)\right)},$$

where $\hat{\delta}_j$ denotes the estimate of the treatment effect and $s(\delta)_j^2$ is a consistent estimator of the variance of δ_j. Since the maximum of the denominator is unrestricted it equals 1 at $\hat{\delta}_j = \delta_j$. Therefore, the Gail and Simon test statistic reduces through the numerator. The Gail and Simon test rejects H_0 from Equation (5.2) if the minimum of $Q^- = \sum_{j=1}^{J} \left(\hat{\delta}_j{}^2 / s(\delta)_j^2\right) I(\delta_j < 0)$ and $Q^+ = \sum_{j=1}^{J} \left(\hat{\delta}_j{}^2 / s(\delta)_j^2\right) I(\delta_j > 0)$ is greater than an appropriate critical value. Where $I(\delta_j > 0) = 1$ if $\delta_j > 0$ and 0 otherwise, and $I(\delta_j < 0) = 1$ if $\delta_j < 0$ and 0 otherwise. They showed that the Q statistic follows a weighted sum of χ^2 distributions. The p-value for the test statistic Q is computed as

$$p_Q = \sum_{j=1}^{J-1} (1 - \chi_j(Q))\, bin(j; n = J-1, p = 0.5),$$

where$\chi_j()$ is the cumulative chi-square distribution function with j degrees of freedom and $bin(j; n, p)$ is the binomial probability function with parameters n and p. Using the Gail and Simon test it is also possible to test the one sided null hypothesis that all treatment effects δ_j are positive (Q^+), or the one sided null hypothesis that all treatment effects are negative (Q^-) [Dmitrienko et al., 2005]. The corresponding p-values for the test statistics are

$$p_{Q^+} = \sum_{j=1}^{J} \left(1 - \chi_j(Q^+)\right) bin(j; n = J, p = 0.5),$$

and

$$p_{Q^-} = \sum_{j=1}^{J} \left(1 - \chi_j(Q^-)\right) bin(j; n = J, p = 0.5).$$

Please note, that the Gail and Simon test statistic uses a standardized sum of squares of the treatment effects. Therefore, it cannot provide any information about the source of a potential reversal treatment effect.

5.1.3. Piantadosi and Gail test

Piantadosi and Gail [1993] proposed a standardized range test to test the null hypothesis of no qualitative interaction by considering the minimum and the maximum of the observed treatment effects over the levels of the second factor. The

null hypothesis is rejected at level α if both

$$\max\left\{\hat{\delta}_j / s(\delta)_j\right\} > r_\alpha \qquad \text{and} \qquad \min\left\{\hat{\delta}_j / s(\delta)_j\right\} < -r_\alpha. \tag{5.3}$$

The critical value r_α is determined by

$$r_\alpha = \Phi^{-1}\left(\frac{1 - (2(1-\alpha)^{\frac{1}{J-1}} - 1)}{2}\right),$$

where Φ corresponds to the standard normal distribution.

Unfortunately, none of the global tests to detect qualitative interactions is appropriate to investigate the source and the amount of a potential qualitative interaction. In the following, the approach developed by Kitsche and Hothorn [2013] is introduced that allows those in depth analysis of qualitative interactions.

5.2. Detecting qualitative interactions using ratios of treatment differences

Kitsche and Hothorn [2013] proposed to use the ratios of treatment differences to detect qualitative interactions. The appropriate formulation of those ratios of treatment differences was presented in Section 3.4 and the related inferential procedures are given in Section 4.4. To illustrate the main concept of Kitsche and Hothorn [2013] Figure 5.1 displays the parameter space of two treatment effects. A treatment effect denotes the difference between two levels of the primary treatment factor, e.g., $\delta_j = \mu_{\text{Drug},j} - \mu_{\text{Placebo},j}$ from the multi-centre clinical trial in Section 2.3.

From Figure 5.1, it is clear that the use of product-type interaction contrasts, as described by Gabriel et al. [1973], is inappropriate for differentiating between quantitative and qualitative interactions because this approach uses the differences between the treatment effects. A straightforward solution to this problem is the usage of the ratios of the treatment effects: if a qualitative interaction is present, then the ratio of the treatment effects receives a negative sign. In contrast, if there is no qualitative interaction present, then the sign of the ratio of the treatment

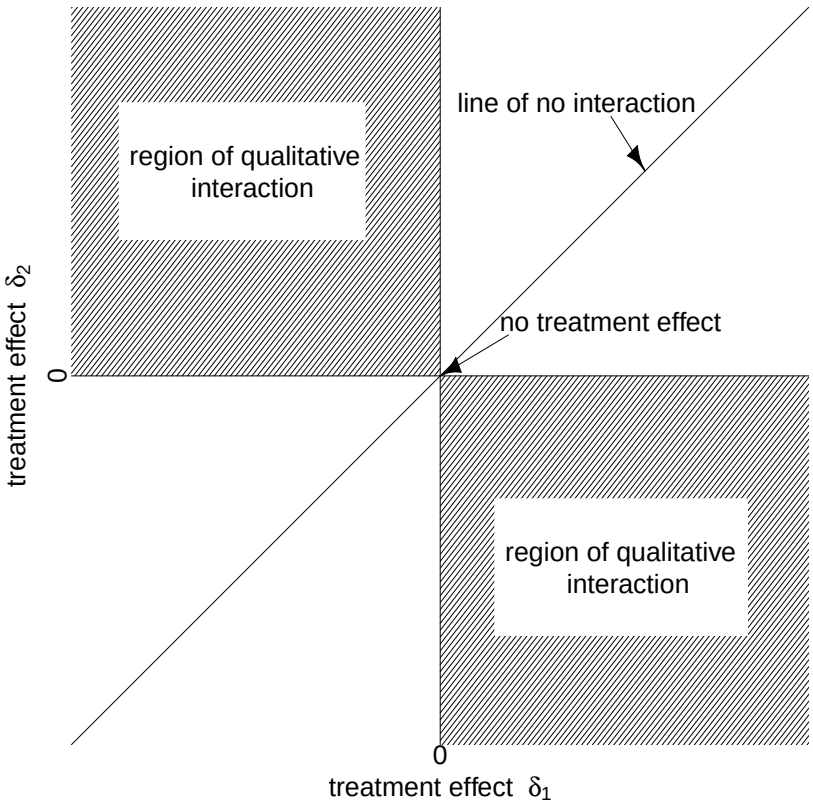

Figure 5.1.: Parameter space of the treatment effects $\boldsymbol{\delta} = \{\delta_1, \delta_2\}$ for two levels of the secondary factor. The shaded regions define the space of qualitative interactions where the signs of δ_1 and δ_2 differ. The diagonal line refers to the set of parameter combinations in which no interaction effect is present, and the origin represents the parameter combination in which no treatment effect exists.

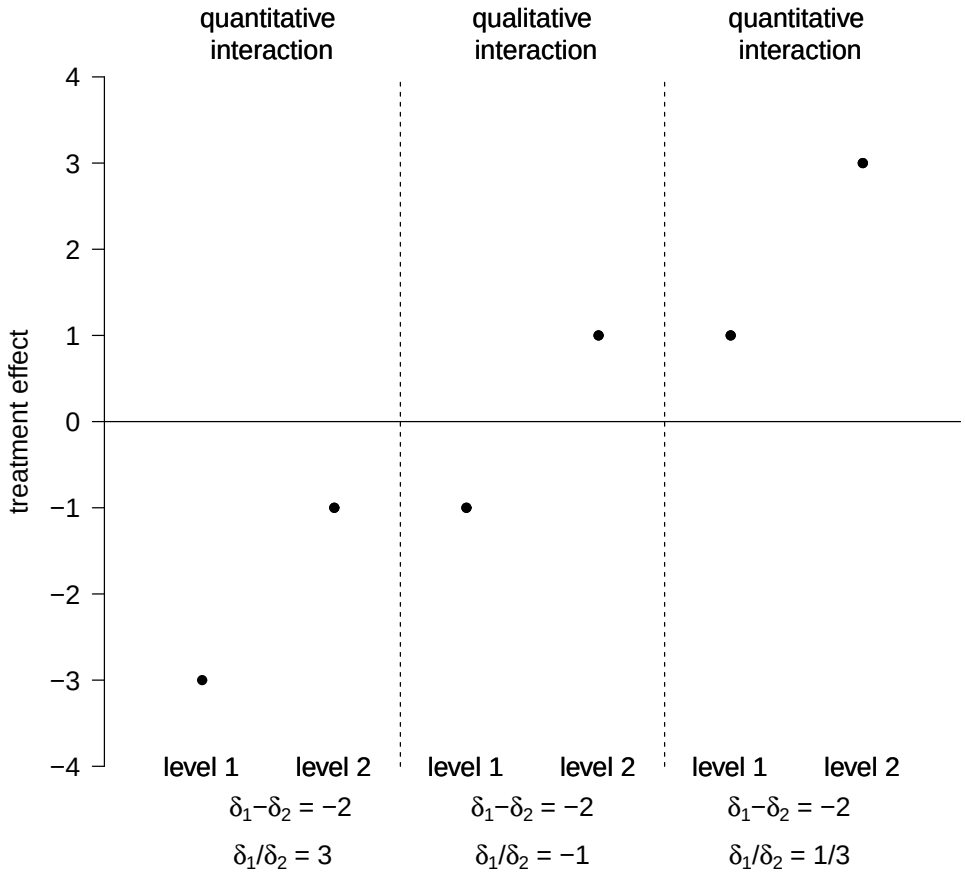

Figure 5.2.: Three scenarios for the treatment effects $\boldsymbol{\delta} = \{\delta_1, \delta_2\}$. The first scenario describes the case of a quantitative interaction in which both treatment effects are negative. The second scenario considers the case in which the treatment effects differ in sign (a qualitative interaction), and the third case represents the situation of a quantitative interaction in which both treatment effects are positive. In all of the scenarios, the difference in the treatment effects is equal.

effects becomes positive. As a consequence, the one-sided version of the global hypothesis in Equation 3.11 is appropriate to detect a qualitative interaction, whereas all elements of the relative margin $\boldsymbol{\omega}$ are set to zero:

$$H_0: \quad \boldsymbol{\gamma} := \frac{\boldsymbol{H\mu}}{\boldsymbol{D\mu}} \geq \mathbf{0}, \tag{5.4}$$

where $\boldsymbol{H}$ and $\boldsymbol{D}$ represent the $M \times L$ numerator and denominator contrast matrix, respectively. Each research hypothesis can now be formulated by the mth row vector of the numerator and denominator contrast matrix, namely $\boldsymbol{h_m}$ and $\boldsymbol{d_m}$:

$$H_0^{(m)}: \qquad \gamma^m := \frac{\boldsymbol{h}_m\boldsymbol{\mu}}{\boldsymbol{d}_m\boldsymbol{\mu}} \geq 0.$$

By means of illustration, consider the three situations in Figure 5.2, where the treatment effects $\boldsymbol{\delta} = \{\delta_1, \delta_2\}$ are displayed for three scenarios. In the first and third scenario, a quantitative interaction is present, whereas in the second situation, a qualitative interaction exists. Nevertheless, in all of the cases, the difference in the treatment effect is equal: hence, no distinction between qualitative and quantitative interactions is possible in these cases by using the difference in the treatment effects. In contrast, using the ratio of the treatment effects allows for a distinction between the quantitative and qualitative interactions. In the second case, the sign of the ratio of the treatment effects is negative and, therefore, a qualitative interaction must be concluded. In contrast, the ratio of the treatment effect in the first and third scenario results in a positive value and, therefore, a quantitative interaction must be inferred.

5.2.1. Geometric representation of qualitative interactions

In this subsection the geometric approach introduced in Subsection 4.5.2 is applied to illustrate the distinction between quantitative and qualitative interactions using the Fieller type confidence intervals for the ratios of treatment effects. Figure 5.3 displays three scenarios (A,B and C) for the parameter space for the treatment effects δ_1 and δ_2 (this corresponds to the parameter space displayed in Figure 5.1). In plot A and B the point (δ_2, δ_1) (red point) lies in the region of qualitative interaction since the treatment effect differ in their sign (see also Figure 5.1). As

defined in Subsection 4.5.2 the corresponding ratio $\gamma = \frac{\delta_2}{\delta_1}$ is the slope of the line defined by the connection of (δ_1, δ_2) to the origin $(0, 0)$. Obviously, the ratio is appropriate to detect a qualitative interaction: if the slope (ratio) is negative, the the sign of the treatment effects differs. Plot C considers the case of no significant qualitative interaction. In this case the confidence ellipse also lies in the quadrant of no qualitative interaction and therefore the upper confidence limit for the ratio $\gamma = \delta_2/\delta_1$ is greater than zero concluding on no significant qualitative interaction.

5.2.2. Characteristics of the test for qualitative interactions

In this subsection the characteristics of the test to detect a qualitative interaction are discussed using the geometric interpretation introduced in Subsection 4.5.2. Figure 5.4 displays the parameter space for two treatment effects δ_1 and δ_2 for $\omega = 1$, $\omega = 0.5$ and $\omega = 0$ defined by the local null hypothesis:

$$H_0^m : \frac{\boldsymbol{h}_m \boldsymbol{\mu}}{\boldsymbol{d}_m \boldsymbol{\mu}} \geq \omega.$$

Note, that the last situation corresponds to the hypothesis of no qualitative interaction. From the plots in Figure 5.4 it is obvious that with an ω approaching to zero the influence of the denominator decreases. In the extreme case of $\omega = 0$ the information provided by the denominator drops out, meaning that the slope (ratio) reduces to $\gamma = \delta_2 = \boldsymbol{h}_m \boldsymbol{\mu}$ and is equal to zero. Using the test statistic in Equation 4.5 and testing for the presence of qualitative interaction by setting $\omega = 0$ results in the test statistic

$$T_m = \frac{\boldsymbol{h}_m \hat{\boldsymbol{\mu}}}{s\sqrt{\boldsymbol{h}_m \boldsymbol{M} \boldsymbol{h}_m^T}}.$$

Obviously, the information provided by the denominator interaction contrast matrix gets lost.

Since the numerator contrast represents the mth treatment effect δ_m this test statistic tests the null hypothesis that the mth treatment effect is smaller than zero. This hypothesis directly corresponds to the global one-sided hypothesis of

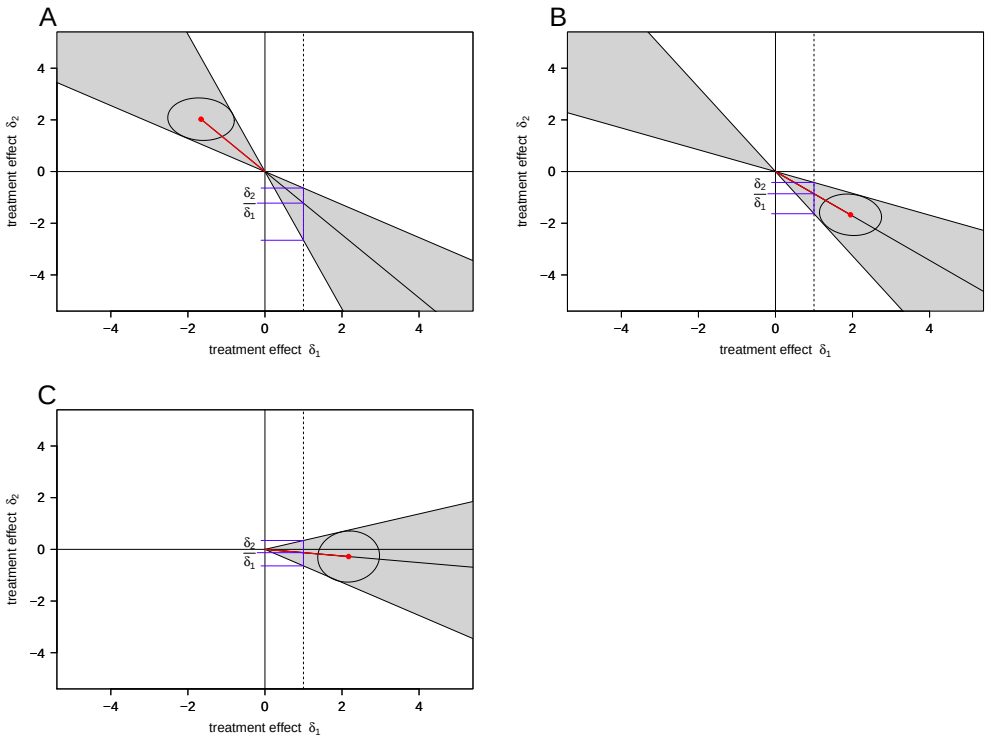

Figure 5.3.: The geometric representation of the Fieller type confidence intervals for $\gamma = \delta_2/\delta_1$ to detect qualitative interactions. The confidence bounds are defined by the rays from the origin that are tangent to the ellipse that defines the 100 $(1 - \alpha)$ ellipse for a combination of the parameters in γ. **A**: Case of significant qualitative interaction: the point (δ_1, δ_2) lies in the upper left quadrant which corresponds to a qualitative interaction. The resulting lower confidence limit for the ratio δ_2/δ_1 at $x = 1$ is smaller than zero. **B**: Case of significant qualitative interaction: the point (δ_1, δ_2) lies in the lower right quadrant which corresponds to a qualitative interaction. The resulting lower confidence limit for the ratio δ_2/δ_1 at $x = 1$ is smaller than zero. **C**: Case of no significant qualitative interaction: the point (δ_1, δ_2) lies in the lower right quadrant which corresponds to a qualitative interaction. Nevertheless, the resulting upper confidence limit for the ratio δ_2/δ_1 at $x = 1$ is greater than zero.

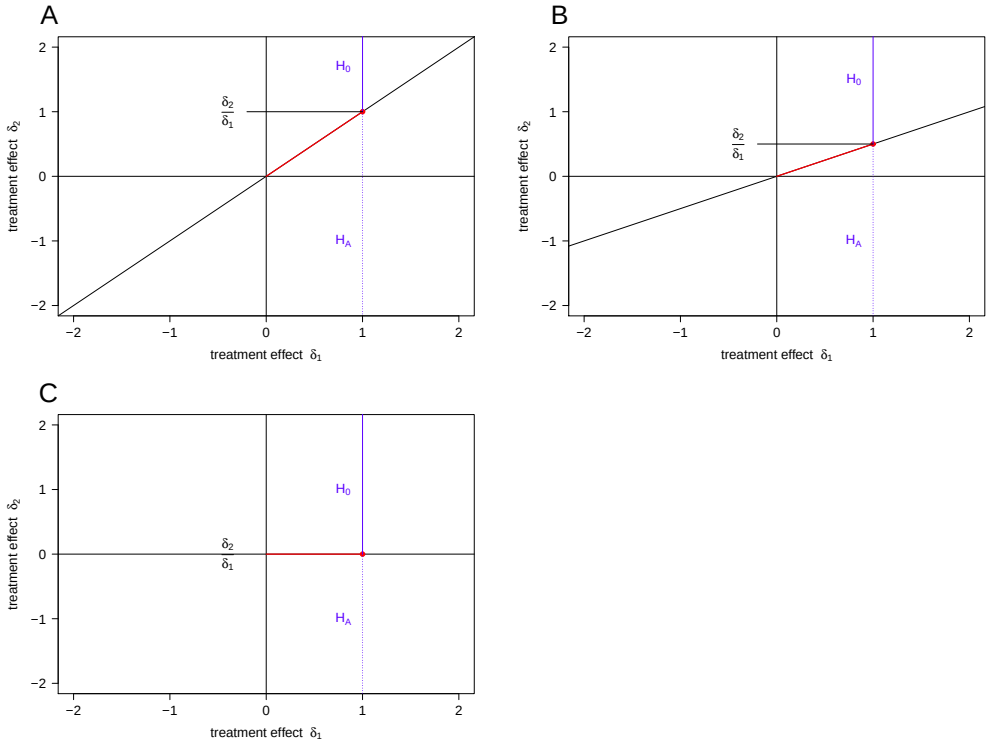

Figure 5.4.: The geometric representation of the ratio $\gamma = \delta_2/\delta_1$. **A**: Scenario to test the null hypothesis $\gamma \geq 1$. **B**: Scenario to test the null hypothesis $\gamma \geq 0.5$. **C**: Scenario to test the null hypothesis $\gamma \geq 0$ of no qualitative interaction.

the Gail and Simon test, that all treatment effects are smaller than zero:

$$H_0 : \boldsymbol{\delta} \in \boldsymbol{O}^+.$$

Therefore, the proposed test to detect a qualitative interaction is only applicable if the direction of the considered treatment effect is a priori known.

Chapter 6.

Consistency assessment

The global tests and the proposed method presented in Section 5 are all constructed to detect a qualitative interaction. Nevertheless, as stated by Wellek [1997]: "*the primary aim of considering qualitative interactions is to establish the homogeneity of the sub-populations with respect to the direction of the treatment effects*", because a positive result of a test on detecting a qualitative interaction is an undesirable outcome. More generally speaking, the detection of no qualitative interaction corresponds to the assessment of consistency of the treatment effect over the predefined subgroups. In the case of no qualitative interaction the consistency is declared if the treatment effects are equal in their sign.

In recent years the assessment of consistency of the treatment effects gained enlarged attention in the field of multi-regional clinical trials. Several regulatory health authorities addressed this issue. For example the Ministry of Health, Labour and Welfare of Japan [2007] proposed that the observed treatment effect for Japanese patients should be at least half of that observed for all patients to accept consistency of the treatment effect. According to Chen et al. [2010] the problem of consistency assessment of the treatment effect in a multi-regional clinical trial conforms to the problem of non-inferiority testing in medical research. Thereby non-inferiority of the treatment effects is declared if the treatment effects are non-inferior in comparison to some competitive treatment effect for a pre-specified irrelevant amount. A meaningful competitive treatment effect is the overall treatment effect of the trial, see, e.g. the Ministry of Health, Labour and Welfare of Japan [2007]. The irrelevant deviation is also denoted as consistency margin. Within this chapter it is demonstrated how to use the ratios of differences

among means and the related inferential methodology presented in Section 4.2 to assess consistency of the treatment effect over pre-defined subgroups.

Hauschke and Kieser [2001] presented an approach to establish non-inferiority of several treatment means compared to a control mean by using the ratio. Dilba et al. [2006b] addressed the problem of power and sample size calculation in non-inferiority trials based on the ratios of treatment means to a control. Their approach is adopted to assess non-inferiority of treatment effects by building the ratios of linear combinations of treatment means. The corresponding parameters of interest γ_m are again formulated as in Section 3.4. Considering a primary treatment factor with two levels the differences of treatment means are again denoted as treatment effects $\delta_j = \mu_{1j} - \mu_{2j}$.

Assessment global consistency The paper from Hauschke and Kieser [2001] addresses the situations to demonstrate global and partial non-inferiority. In the situation of global non-inferiority it must be shown that the treatment effect is consistent for all pre-defined subgroups. Hence, the global null and alternative hypotheses can be formulated as

$$H_0 : \quad \bigcup_{m=1}^{M} H_0^m \quad \text{versus} \quad H_A : \quad \bigcap_{m=1}^{M} H_A^m \tag{6.1}$$

where the local hypotheses are given by

$$H_0^m : \frac{\boldsymbol{h}_m \boldsymbol{\mu}}{\boldsymbol{d}_m \boldsymbol{\mu}} \leq \omega \qquad \text{versus} \qquad H_A^m : \frac{\boldsymbol{h}_m \boldsymbol{\mu}}{\boldsymbol{d}_m \boldsymbol{\mu}} > \omega.$$

If the local null hypothesis H_0^m is rejected, a consistency of the treatment effect given the consistency margin ω is inferred. Figure 6.1 displays the parameter space to assess global non-inferiority for two ratios of treatment effects. The parameters of interest are defined as the ratio of one treatment effect to the overall treatment effect, $\gamma_j = \delta_j / \bar{\delta}$. The inconsistency margin ω is the amount of the acceptable relative change of the treatment effect from the jth subgroup in comparison to the overall treatment effect. Please note, that a consistency margin of $\omega = 0$ corresponds to a distinction between quantitative and qualitative interaction.

The analysis of the problem in Equation 6.1 is performed by applying the

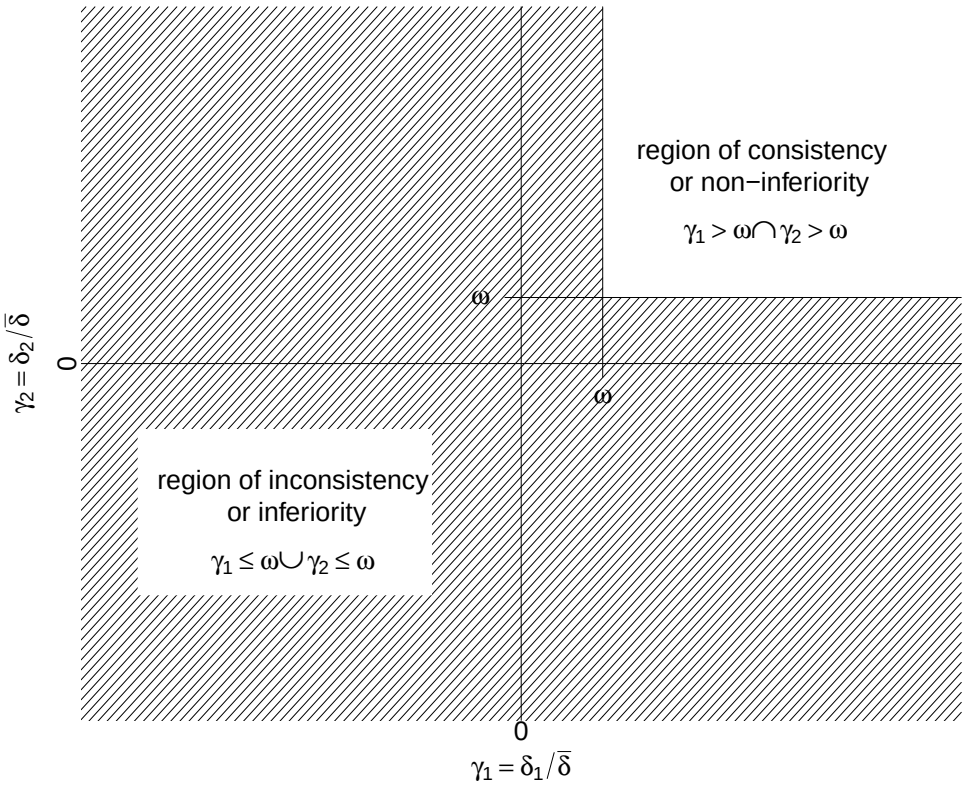

Figure 6.1.: Parameter space for the parameters of interest γ_m for m=2 research hypotheses defined as the ratio of the treatment effect δ_j to the overall treatment effect $\bar{\delta}$. The shaded region defines the region under the null hypothesis for the intersection-union principle to assess global consistency. The null hypothesis in Eq. 6.1 is rejected if and only if γ_1 and γ_2 are greater than the pre-defined consistency margin ω.

intersection-union principle. This means, that the global null hypothesis H_0 is rejected if and only if each of its local null hypothesis H_0^m is rejected.

Within the intersection-union testing principle no multiplicity adjustment is needed [Dmitrienko and D'Agostino, 2013]. Therefore, each local null hypothesis H_0^m is tested at the full α level. In this case each local test statistic T_m in Equation 4.9 follows a t-distribution with ν degrees of freedom. The global null hypothesis H_0 is rejected if the minimum of the vector of test statistics $\boldsymbol{T} = (T_1, \ldots, T_M)^T$ is greater than the $1 - \alpha$ quantile of the t-distribution. The associated unadjusted p-values can be calculated as $p_m = P(t \geq T_m)$, where the variable t follows a

t-distribution with ν degrees of freedom. In addition, the corresponding one-sided marginal confidence intervals can be applied. In this case, the global null hypothesis H_0 is reject if all lower confidence limits are greater than the pre-defined consistency margin ω. Please note, that a rejection of the global null hypothesis does not allow any conclusions on the local hypotheses. If non-inferiority cannot be inferred for all local hypotheses, the question of which sub-groups are non-inferior cannot be answered [Hasler and Hothorn, 2013].

Assessment local consistency The formulation of the global null hypotheses as an intersection-union test provides only information for the global hypothesis. Nevertheless, in most circumstances interest is in inferences on the local hypotheses, e.g., if interest is in consistency assessment of the treatment effect for certain regions or countries in a multi-regional clinical trial. According to Chen et al. [2010], such a local assessment of consistency is often requested by local regulatory agencies to support registration in multi-regional clinical trials. Consider for example the Ministry of Health, Labour and Welfare of Japan [2007].

Therefore the global hypotheses are now defined in the context of an union-intersection testing problem:

$$H_0 : \quad \bigcap_{m=1}^{M} H_0^m \quad \text{versus} \quad H_A : \quad \bigcup_{m=1}^{M} H_A^m. \tag{6.2}$$

The global null hypothesis is rejected, if at least one of the m local hypotheses is rejected. In the union-intersection testing approach a multiplicity adjustment is needed to protect the overall type I error rate at the nominal level α [Dmitrienko and D'Agostino, 2013]. To adjust for multiplicity in the context of local consistency assessment for several sub-groups the multiple comparison procedure presented in Section 4.4 can be applied.

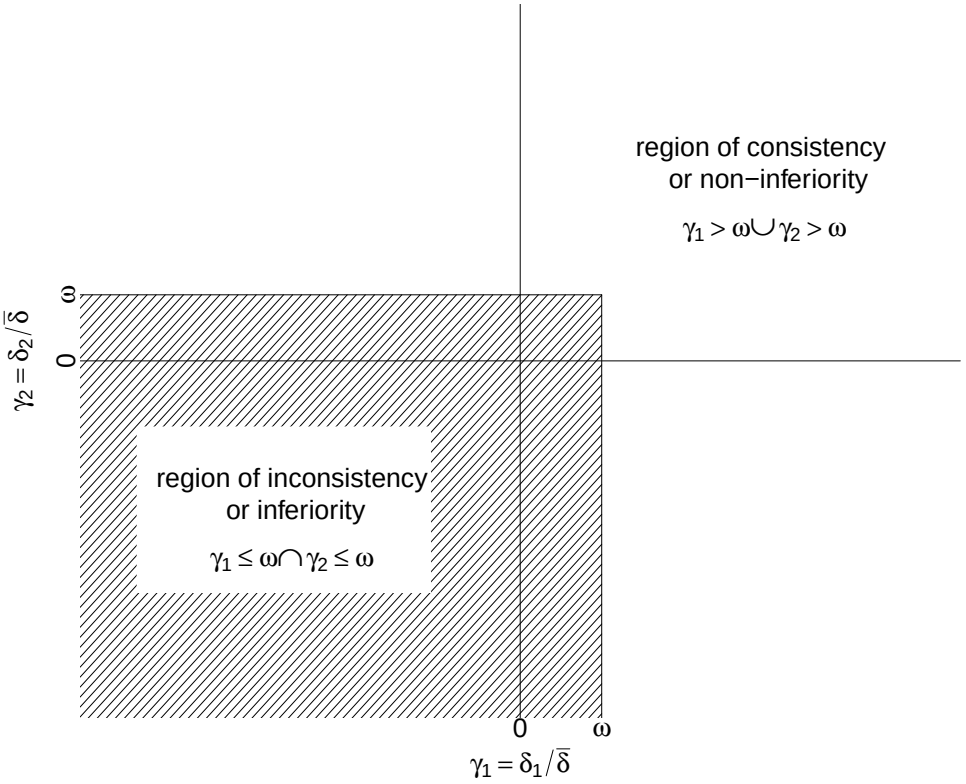

Figure 6.2.: Parameter space for the parameters of interest γ_m for m=2 research hypotheses defined as the ratio of the treatment effect δ_j through the overall treatment effect $\bar{\delta}$. The shaded region defines the region under the null hypothesis for the union-intersection principle to assess local consistency. The null hypothesis is rejected if at least one γ_m is greater than the pre-defined consistency margin ω.

Chapter 7.

Binomial data

Up to this point, is was assumed that the primary response is a normally distributed outcome variable. Nevertheless, in biological and biomedical research the outcome of interest is often an "event" with the data taking a binary form commonly denoted as success or failure. Therefore, the goal of this chapter is to present a methodology that is suitable for the analysis of statistical interactions in the presence of a binary outcome measure. Furthermore, the focus here is to detect qualitative interaction, because those interactions are of greater importance than quantitative interactions [Peto, 1982, Baker, 1988]

7.1. The model

Again, a completely randomized design including one treatment factor and one pre-specified stratification factor is supposed. Let I be the number of groups of the first factor (with index $i = 1, \ldots, I$), e.g., representing the treatment groups. Furthermore, let J be the number of groups of the second factor (with index $j = 1, \ldots, J$), e.g., representing levels of regions, centres, environments, *etc.* The number of observations for each factor combination is permitted to vary and is denoted as n_{ij}. The total number of observations in the study is given by $N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. The primary endpoint is a binary outcome Y represented by 1 and 0, with the generic labels success and failure. The total number of successes in each factor combination is given by y_{ij}. Representative contingency tables for this study design are given in Table 7.1, where the dot notation means the sum over the corresponding factor.

Table 7.1.: Contingency tables for the supposed study design including one treatment factor ($i = 1, \ldots, I$) and one pre-specified stratification factor ($j = 1, \ldots, J$). Here the dot notation denotes the sum over the corresponding factor.

	Outcome		
Treatment	Success	Failure	Total
1	y_{11}	$n_{11} - y_{11}$	n_{11}
2	y_{21}	$n_{21} - y_{21}$	n_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	y_{I1}	$n_{I1} - y_{I1}$	n_{I1}
Total	$y_{.1}$	$n_{.1} - y_{.1}$	$n_{.1}$

$\ldots$

	Outcome		
Treatment	Success	Failure	Total
1	y_{1J}	$n_{1J} - y_{1J}$	n_{1J}
2	y_{2J}	$n_{2J} - y_{2J}$	n_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	y_{IJ}	$n_{IJ} - y_{IJ}$	n_{IJ}
Total	$y_{.J}$	$n_{.J} - y_{.J}$	$n_{.J}$

Further on, it is assumed that y_{ij} follows a binomial distribution with parameters n_{ij} and π_{ij}, denoted by $bin(n_{ij}, \pi_{ij})$. The parameter π_{ij} corresponds to the success probability of the ith treatment in the jth subset:

$$\pi_{ij} = P(Y = 1 | I = i, J = j). \tag{7.1}$$

The maximum likelihood estimator for the sample proportions is given by $\hat{\pi}_{ij} = y_{ij}/n_{ij}$ and its standard error by $\hat{\sigma}(\pi_{ij}) = \sqrt{\pi_{ij}(1 - \pi_{ij})/n_{ij}}$ [Agresti, 2013]. The vector of success probabilities for each factor combination is defined by $\boldsymbol{\pi} = (\pi_{11}, \pi_{21}, \ldots, \pi_{1J}, \pi_{2J})^T$, where the elements of $\boldsymbol{\pi}$ are primarily ordered according to the stratification factor and within the levels of the stratification factor according to the treatment factor. For the sake of the presented method, the vector $\boldsymbol{\pi}$ is given the new index l yielding to $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^T$, where $L = I \cdot J$ is the number of treatment-by-subset combinations. For illustrative purposes and without loss of generality, a study with one control and one treatment group ($I = 2$) is assumed. Furthermore, a treatment effect is defined as the difference of success probabilities between the two treatment groups, also known as risk difference, $\beta_j = \pi_{1j} - \pi_{2j}$, with their standard errors $\hat{\sigma}(\beta_j) = \sqrt{\frac{\pi_{1j}(1-\pi_{1j})}{n_{1j}} + \frac{\pi_{2j}(1-\pi_{2j})}{n_{2j}}}$ [Agresti, 2013]. In the following, it is assumed that the sample size is large enough to suppose asymptotic normality of the maximum likelihood estimator.

7.2. Inference for ratios of risk differences

Several global tests for the null hypothesis of the homogeneity of the treatment effect between strata for binary response variables are available, e.g., the Breslow-Day test [Breslow and Day, 1994] or a likelihood-ratio test [Agresti and Hartzel, 2000, Agresti, 2013]. Nevertheless, these tests provide only global inference and do not distinguish between quantitative and qualitative interactions. Dmitrienko et al. [2005] also used the Gail and Simon test to test for qualitative interaction in the presence of a binary response variable. In this case the estimates of the risk differences β_j and its standard error $\hat{\sigma}(\beta_j)$ are used to calculate the test statistic $Q = \min(Q^-, Q^+)$ (see, Section 5.1.2). Nonetheless, the Gail and Simon test provides only global inference concerning a qualitative interaction.

Since the objective is to detect the source of a potential qualitative interaction the methodological concept developed by Kitsche and Hothorn [2013] to detect qualitative interactions for normally distributed outcome measures is used here. Therefore, the research hypotheses are formulated as the ratios of linear combinations of binomial proportions:

$$\gamma_m = \frac{\sum_{l=1}^{L} h_{lm} \cdot \pi_i}{\sum_{l=1}^{L} d_{lm} \cdot \pi_i} = \frac{\boldsymbol{h}_m \boldsymbol{\pi}}{\boldsymbol{d}_m \boldsymbol{\pi}}, \qquad m = 1, \ldots, M, \tag{7.2}$$

where $\boldsymbol{h}_m$ and $\boldsymbol{d}_m$ represent, respectively, the mth numerator and denominator contrasts, which define linear combinations of proportions. Again, the vectors of the linear combinations are stored in the $M \times L$ numerator and denominator interaction contrast matrices $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_M)^T$ and $\boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M)^T$. The numerator and denominator interaction contrast matrices are constructed according to the scheme in Section 3.4. The parameters of interest are interpretable as the ratios of user defined risk differences. The global null hypothesis of no qualitative interaction and the corresponding alternative can be formulated as:

$$H_0 : (\gamma_1 \geq \omega \cap \gamma_2 \geq \omega \cap \ldots \cap \gamma_M \geq \omega) \tag{7.3}$$

and

$$H_A : (\gamma_1 < \omega \cup \gamma_2 < \omega \cup \ldots \cup \gamma_M < \omega). \tag{7.4}$$

To test for the presence of qualitative interaction, ω equals 0.

Test for ratios of risk differences Djira et al. [2010] discussed methods for simultaneous inference of multiple ratios of binomial proportions. In their paper Djira et al. [2010] considered the case of comparing individual binomial proportions to the pooled population proportion, assuming that their unbiased estimators are asymptotically normally distributed. The ratios of linear combinations of binomial proportions in Equation 7.2 can be reformulated as the linear form $L_m = (\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{\pi}}$. Under H_0 this linear form is approximately normally distributed with mean zero and variance [Djira et al., 2010]:

$$s^2_{L_m} = (\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)^T.$$

Using this linear form and dividing it by the corresponding standard error leads to the test statistic:

$$Z_m = \frac{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{\pi}}}{\sqrt{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)^T}}, \tag{7.5}$$

where $\boldsymbol{M}$ is the diagonal matrix containing the reciprocals of the sample sizes n_{ij} and $\hat{\boldsymbol{V}}$ is a diagonal matrix containing the estimated group variances. According to Djira et al. [2010] the vector of test statistics $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^T$ approximately follows a m-variate normal-distribution with a zero vector of means and a correlation matrix $\hat{\boldsymbol{R}} = [\hat{\rho}_{mm'}]$, where each element of $\hat{\boldsymbol{R}}$ can be described by

$$\begin{aligned} \hat{\rho}_{mm'} &= \mathrm{Corr}(Z_m, Z_{m'}) \\ &= \frac{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_{m'} - \omega \boldsymbol{d}_{m'})^T}{\sqrt{(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_m - \omega \boldsymbol{d}_m)^T}\sqrt{(\boldsymbol{h}_{m'} - \omega \boldsymbol{d}_{m'})\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_{m'} - \omega \boldsymbol{d}_{m'})^T}}, \end{aligned} \tag{7.6}$$

where $m \neq m'$. The global null hypothesis in Equation 7.3 is rejected if $Z_m < c_{\alpha,\hat{R}}$ for at least one m. The critical point $c_{\alpha,\hat{R}}$ is the α-level equi-coordinate percentage point of an m-variate normal-distribution with the correlation matrix

$\hat{\boldsymbol{R}}$. The associated, one-tailed, adjusted p-values can be calculated as $p_m = P(z \leq Z_m)$, where the variable z follows the multivariate normal-distribution with the correlation matrix $\hat{\boldsymbol{R}}$.

Simultaneous confidence intervals for ratios of risk differences In contrast to adjusted p-values, simultaneous confidence intervals for the parameters γ_m would provide information on the amount of qualitative interactions and could therefore be used to assess their clinical relevance. The approach published by Djira et al. [2010] to construct simultaneous confidence intervals for multiple ratios of binomial proportions is adopted here. The corresponding test statistics are now defined as

$$Z_m(\gamma_m) = \frac{(\boldsymbol{h}_m - \gamma_m \boldsymbol{d}_m)\hat{\boldsymbol{\pi}}}{\sqrt{(\boldsymbol{h}_m - \gamma_m \boldsymbol{d}_m)\hat{\boldsymbol{V}}\boldsymbol{M}(\boldsymbol{h}_m - \gamma_m \boldsymbol{d}_m)^T}}. \tag{7.7}$$

According to Djira et al. [2010] the simultaneous confidence intervals for γ_m are available by solving the equation $Z_m(\gamma_m) = q$ for some specified quantile q. Djira et al. [2010] proposed to use q as an equi-coordinate percentage point of the joint distribution of the vector of test statistics $\boldsymbol{Z}(\gamma) = (Z_1(\gamma_1), \ldots, Z_M(\gamma_M))^T$. The correlations between the test statistics $\boldsymbol{Z}(\gamma)$ are similar to those defined in Equation 7.6, except that ω is replaced by γ_m and $\gamma_{m'}$. The correlations between two contrasts depend on the user-defined contrasts $\boldsymbol{h}_m$ and $\boldsymbol{d}_m$, the sample sizes n_{ij}, the estimated group variances and the unknown ratios $\boldsymbol{\gamma}$. This paper considers a plug-in approach proposed by Djira et al. [2010], where the unknown ratios in $\hat{\boldsymbol{R}}(\gamma_1, \ldots, \gamma_M)$ are replaced by its maximum likelihood estimators. The null hypothesis of no qualitative interaction in Equation 7.3 is rejected if the upper confidence limit is smaller than the margin $\omega = 0$.
The function tools for the calculations of simultaneous confidence intervals for the ratios of user-defined linear combinations of proportions are available in the add-on package `mratios` Dilba et al. [2012] (function `gsci.ratio()`) for the statistical software package `R` [R Core Team, 2013].

Chapter 8.

Monte Carlo Simulations

In this chapter the behaviour of the methodology to detect qualitative interactions by using the ratios of treatment differences is analysed via Monte Carlo simulations. The performance of the proposed method is determined in terms of the empirical power to detect a qualitative interaction. Furthermore, Monte Carlo simulations for different parameter settings were conducted to investigate the adequacy of the asymptotic approximation of the proposed confidence intervals for binary response variables.

8.1. Power and coverage probability calculations

In the first instance, the empirical power for different parameter settings for the proposed method was calculated. The empirical power was computed as the rate of rejected null hypotheses out of the simulated data sets. For the proposed method, the global null hypothesis was rejected if any of the multiplicity adjusted p-values p_m, corresponding to the mth hypothesis, was smaller than the pre-specified significance level $\alpha = 0.05$:

$$P(p_m \leq 0.05; \exists m = 1, \ldots, M), \tag{8.1}$$

where p_m denotes the multiplicity adjusted p-value corresponding to the mth hypothesis. Please note, that this power definition is commonly known as any-pair power.

Second, the simultaneous coverage probabilities of the one-sided confidence in-

tervals for the binary response variables were investigated to analyse their asymptotic approximation. Within here, the coverage probability was defined as the probability that each true value γ_m, for $m = 1, \ldots, M$ ratios of treatment effects of interest, is less than or equal to the corresponding upper bound of the simultaneous confidence intervals (U_m). In mathematical notation, the coverage probability is defined as

$$P(\gamma_m \in [-\infty, U_m], \forall m = 1, \ldots, M). \tag{8.2}$$

8.2. Simulations for normally distributed outcome variables

8.2.1. Setup

The simulation studies are based on the design from the multi-centre clinical trial example of Section 2.3 with a continuous response variable $\boldsymbol{x}$ that follows a normal distribution with mean μ_l and a common standard deviation σ, $\boldsymbol{x} \sim N(\mu_l, \sigma^2)$. For the simulation study, the number of levels of the primary treatment factor was set to 2, e.g., representing an active treatment and a placebo group. Furthermore, a total number of 10 levels of the secondary factor were considered, that represent, e.g., centres, subgroups or regions. For each simulated data set, the vector of true cell means was set to $\boldsymbol{\mu}^T = (10, 10 + \varphi, 10, 10 + \varphi, \ldots, 10, 10 + \varphi)$ and the sample size for each group to 30. The variance was set to $\sigma^2 = 5$ for each sample. For several choices of $\boldsymbol{\mu}$, the power was estimated empirically using Equation 8.1. For each parameter setting, 10,000 data sets were simulated. The shift parameter φ was set from 0 to 2 with increments 0.1 to simulate an increasing treatment effect. To examine the empirical power, the sign of the shift parameter φ was reversed for one or several levels of the secondary factor. For example, to simulate one reverse treatment effect compared to the remaining treatment effects, the vector of cell means was set to $\boldsymbol{\mu}^T = (10, 10 - \varphi, 10, 10 + \varphi, \ldots, 10, 10 + \varphi)$. Furthermore, the situations in which the treatment effect of 2, 3, 4 and 5 of 10 levels of the secondary factor differs in its sign compared to the remaining treatment effects was simulated. To compare the presented method to detect qualitative interactions

for normally distributed outcome variables, the tests of Gail and Simon [1985], Piantadosi and Gail [1993] and Azzalini and Cox [1984] were used as reference methods (see Section 5.1). All of these methods test the global null hypothesis of no qualitative interaction. The null hypothesis of the Gail and Simon, Piantadsi and Gail and Azzalini and Cox test are based on the treatment differences δ_j, as in Equation 5.1, 5.2 and 5.3. In comparison, the global null hypothesis of the proposed method is based on the simultaneous assessment of user defined ratios of the treatment differences, as in Equation 7.3. To detect the levels of the secondary factor with reverse treatment effect the ratio of each treatment effect to the overall treatment effect $\gamma_l = \delta_l/\bar{\delta}$ was used as the parameter of interest in Equation 7.3. The corresponding numerator and denominator contrast matrices are given in Equation 8.3 and 8.4.

$$\boldsymbol{C}_{\text{Interaction}}^{\text{Numerator}} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \tag{8.3}$$

$$\boldsymbol{C}_{\text{Interaction}}^{\text{Denominator}} = \begin{pmatrix} -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 & -0.1 & 0.1 \end{pmatrix} \tag{8.4}$$

8.2.2. Results

The empirical power for the mentioned parameter settings was calculated as the rate of rejected null hypotheses out of 10,000 simulated data sets. For the proposed method, the global null hypothesis was rejected if any of the local null hypotheses was rejected. The results of the Monte Carlo simulations are displayed in Figure 8.1, where the different plots correspond to the different numbers of reverse treatment effects. With an increasing shift parameter φ, the empirical power of all of the four considered methods increases. Furthermore, it is obvious that the empirical power of all methods increases with an increasing number of levels of the secondary factor where the treatment effect differs in its sign. In the cases of one or two reverse treatment effects, the proposed method and the Piantadosi and Gail test show a similar empirical power. In contrast, the performance of the Gail and Simon test is inferior to these methods if the treatment effect of one level of the secondary factor is reverse compared to the remaining treatment effects. These differences between the Gail and Simon test and the Piantadosi and Gail test were also observed by the comparative study of Piantadosi and Gail [1993]. Nevertheless, with an increasing number of reverse treatment effects, the Gail and Simon test is more powerful than the proposed method, especially in cases of high treatment effects. For all that, it must be noted that the latter scenarios, which represent situations of high proportions of reverse treatment effects, are rare in practice. In all of the considered cases, the Azzalini and Cox test is less powerful for detecting a qualitative interaction compared with the remaining methods.
Please note, that the major advantage of the proposed method is its ability to detect the source and the amount of a qualitative interaction, rather than its gain in power in special situations.

8.3. Simulations for binary response variables

8.3.1. Setup

For this simulation study again 10 levels of the secondary factor and two levels of the treatment factor were considered. For each simulation setting, the vector

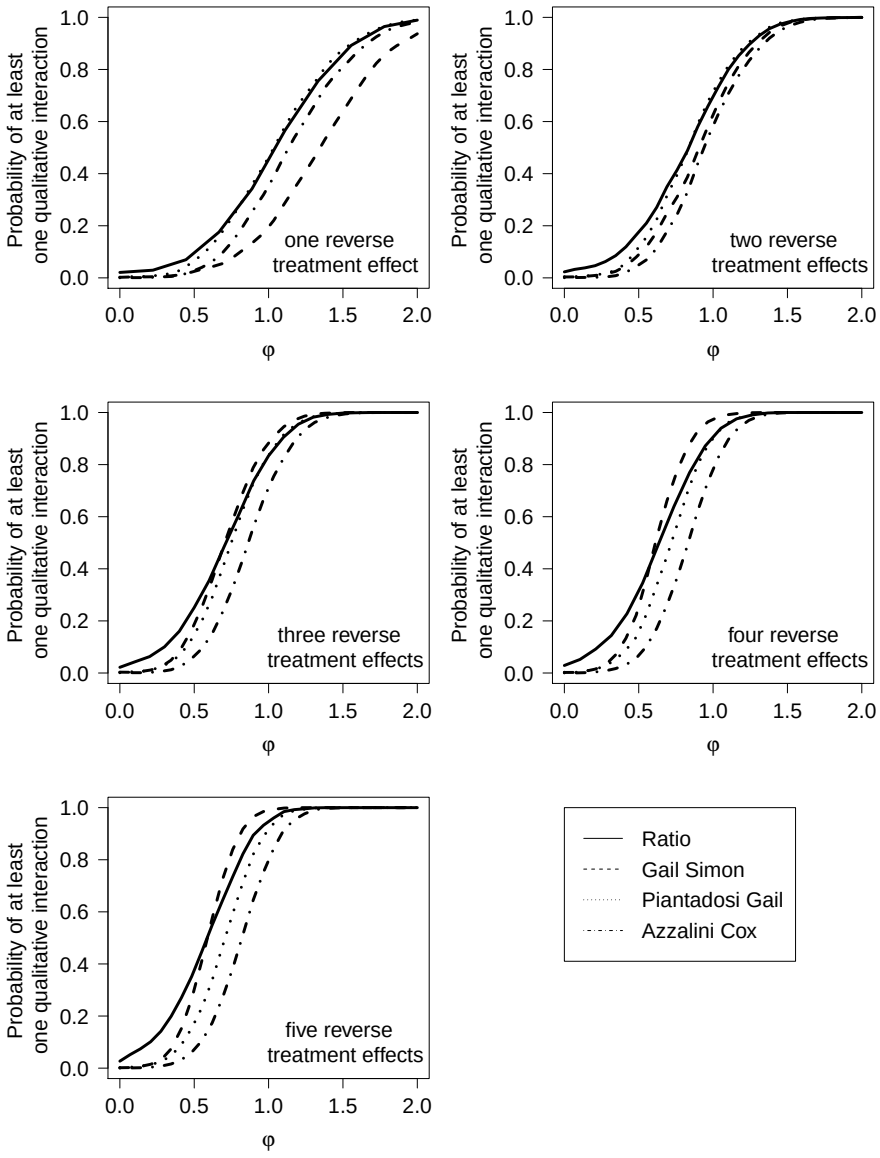

Figure 8.1.: Probability of detecting at least one qualitative interaction in a trial with 10 levels of the secondary factor for an increasing effect size φ. The different plots represent changing numbers of levels of the secondary factor with reverse treatment effect. Dashed lines represent the empirical power for the Gail and Simon test, solid lines represent the proposed method, dotted lines represent the Piantadosi and Gail test and dash-dotted lines represent the Azzalini and Cox test.

of true binomial proportions $\boldsymbol{\pi}$ and the corresponding vector of sample sizes $\boldsymbol{n}$ were defined. For the analysis of the simultaneous coverage probabilities of the one-sided confidence intervals for the parameters γ_m, the proportions in the first subset were set as $\pi_{11} = 0.2 + \delta$ and $\pi_{12} = 0.1$, whereas the proportions in the remaining subsets were set as $\pi_{i=1,j\neq 1} = 0.1$ and $\pi_{i=2,j\neq 1} = 0.2 + \delta$. This design corresponds to a qualitative interaction due to subset one. The shift δ was set to vary between 0 and 0.5 by increments of size 0.01 to increase the treatment effect in the subsets. Furthermore, the sample size for each treatment-by-subset combination was varied: i) $n_{ij} = 40$ for all i and j, ii) $n_{ij} = 30$ for all i and j, iii) $n_{ij} = 20$ for all i and j.

For the power comparisons, the number of reversal subsets were alternated, whereas the number of observations per factor combination was set to $n_{ij} = 40$. The success probabilities for the treatment groups were set to $\pi_{1j} = 0.3$ and $\pi_{2j} = 0.5$, resulting in a treatment effect of $\beta_j = 0.2$. To simulate a qualitative interaction, the treatment effect for a specified number of subsets was inverted by an increasing amount δ ($\pi_{1j} = 0.5 - \delta$ and $\pi_{2j} = 0.5$, with $\delta = 0, \ldots, 0.45$). The number of reversal subsets was set to 1,2 and 3. For each parameter setting, 10,000 data sets were simulated. Again, the numerator and denominator interaction contrast matrices from Equation 8.3 and 8.4 were used to define the parameters of interest γ_m.

8.3.2. Results

Figure 8.2 shows the coverage probability, as defined in Equation 8.2, versus the shift parameter δ for the proposed simultaneous confidence intervals for the parameters γ_m. The confidence level was set to 95% and the number of observations for each treatment-by-subset combination was specified by $n_{ij} = 20, 30$ and 40. The under-coverage of the proposed intervals decreases as the parameter δ increases. Furthermore, the coverage probability converges to the nominal level of 95% with an increasing number of observations per treatment-by-subset combination. These results were expected, since the proposed confidence intervals are based on large sample approximations. Therefore, the SCIs perform well in situations of moderate to high sample sizes.

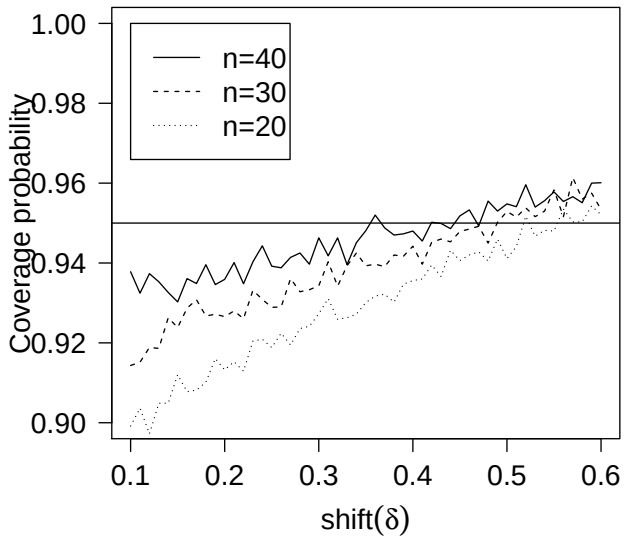

Figure 8.2.: The coverage probability for the simultaneous confidence intervals for the ratios of risk differences γ_m. The horizontal black line marks the nominal level of 95%. An increasing shift parameter δ corresponds to the increasing qualitative interaction. The number of observations for each treatment-by-subset combination was set to $n_{ij} = 20, 30$ and 40 (pointed, dashed and solid line, respectively).

Figure 8.3 presents the empirical power of the proposed method and of the Gail and Simon test to detect a qualitative interaction against the shift parameter δ. The plots show several scenarios: different numbers of subsets with a reversal treatment effect. As expected, for both methods the empirical power increases with an increasing amount of qualitative interaction (increasing δ). Furthermore, the empirical power of both methods to detect a qualitative interaction increases as the number of subsets with a reversal treatment effect increases. In the case of one reversal treatment effect, the proposed method is more powerful than the Gail and Simon test (up to 25%). With an increasing amount of reversal treatment effects the two methods under consideration perform similar. This behaviour was also observed in the simulation study of the normally distributed outcome variables. On balance, it is recommended to use the proposed method since it is at least as powerful as the Gail and Simon test and provides additional information on the source and the amount of the qualitative interaction.

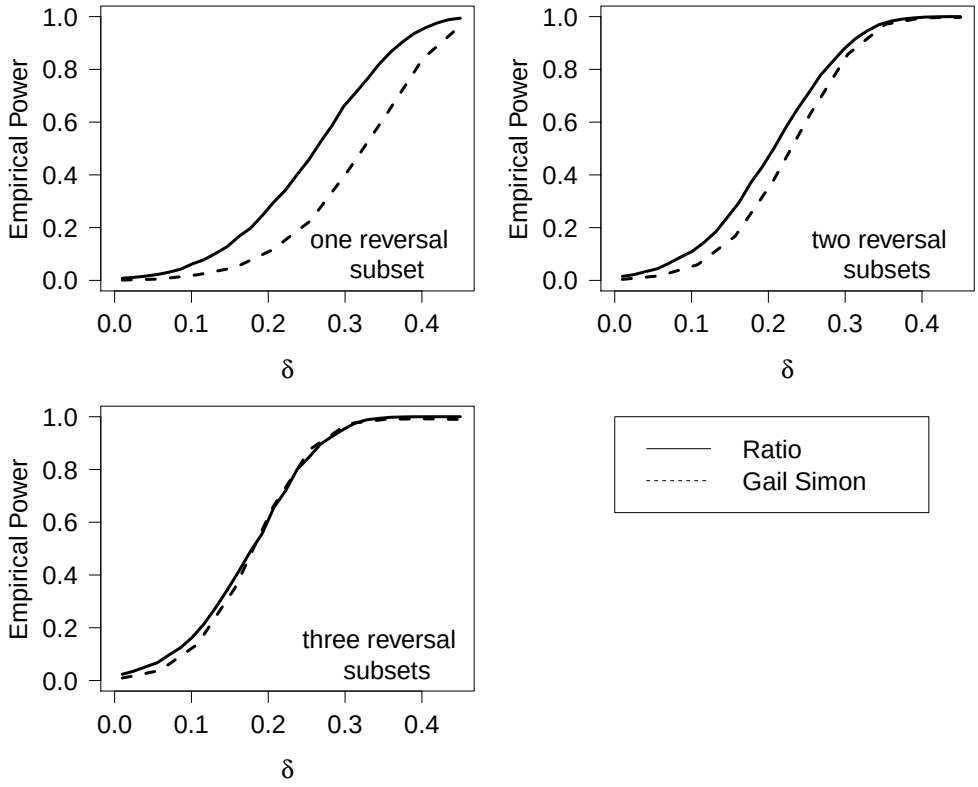

Figure 8.3.: The probability of detecting at least one qualitative interaction among 10 subsets of patients. Solid lines represent the proposed approach (Ratio) dashed lines represent the Gail and Simon test (Gail Simon). Different columns reflect increasing numbers of subsets with reversal treatment effects (1, 2 and 3).

Chapter 9.

Examples re-analysed

Within this chapter the four data examples presented in Chapter 2 are analysed using the methods proposed in the previous chapters for the assessment of potential statistical interactions.

9.1. Bush beans data set

From the two-factorial ANOVA in Table 2.1 a significant spacing-by-variety interaction was detected. In the first instance, the goal is to infer if the spacing effect on mean yield is different between the two growth types. In addition, the evaluation is focused on the detection of a potential variation of the spacing effect in mean yield between the varieties within each growth type. The corresponding product-type interaction contrast matrix to analyse this research hypotheses via MCTs and SCIs is given in Equation 3.5.

Figure 9.1 displays the two-sided simultaneous confidence intervals for the user defined interaction contrasts and the corresponding multiplicity adjusted p-values. For a single interaction contrast the null hypothesis that the user defined contrast is zero is rejected if the confidence interval does not include the value zero. From the top three confidence intervals in Figure 9.1 one can infer that the spacing effect is significantly different between the two growth types on a global level $\alpha = 0.05$. Exemplarily considering the second confidence interval in Figure 9.1 it is concluded that the spacing effect from 20 to 60 cm on the mean yield is at least about 12 kg/plot higher for the bushy varieties than for the tall varieties. Furthermore, it could be concluded that the difference between the two growth

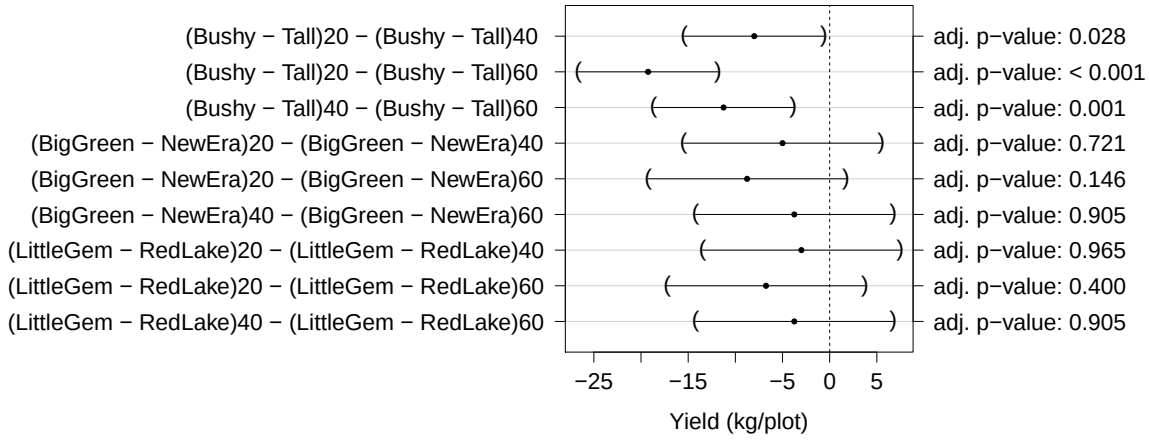

Figure 9.1.: Simultaneous 95% confidence intervals for user defined interaction contrasts as specified in Equation 3.5. Dots denote the estimates for the comparison of interest and vertical bars the lower and upper limit of the two-sided confidence intervals. Adjusted p-values are listed on the right-hand side of each confidence interval.

habits increases with increasing row spacings. In contrast to the comparison of the different spacing effects between the two growth habits, the spacing effect is not different between the varieties of the same growth habit. In summary it can be said, that the signification spacing-by-variety interaction detected by the ANOVA F-test is based on the interaction between the two growth habits and spacing.

Please note, that there are two alternative strategies for the analysis of this data set: (i) an alternative ANOVA, where the hierarchical structure between the growth type and the varieties within the growth type is taken into account (ii) a linear regression on the spacing factor and a subsequent comparison among the regression slopes between varieties.

9.2. Lettuce data set

From the ANOVA in Table 2.2 a significant phosphate-by-soil interaction was detected. The goal is now to investigate the differences of the fertilizer effects between the soil types, where the fertilizer effects of interest are restricted to the differences of each fertilizer to the control group. The appropriate product-type

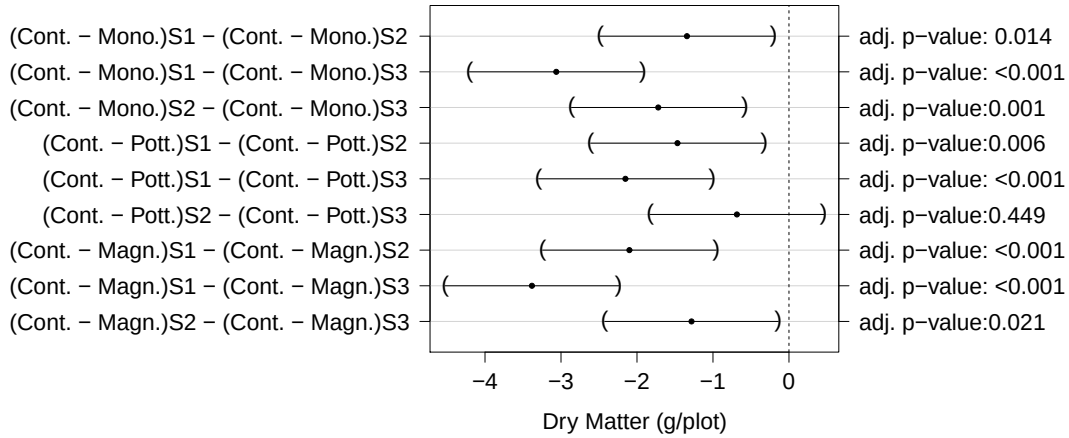

Figure 9.2.: Simultaneous 95% confidence intervals for user defined interaction contrasts as specified in Table 3.7. Dots denote the estimates for the comparison of interest and vertical bars the lower and upper limit of the two-sided confidence intervals. Adjusted p-values are listed on the right-hand side of each confidence interval.

interaction contrasts are given in Equation 3.7. Applying the method presented in Section 4.3 one gets the simultaneous confidence intervals for the nine interaction contrasts of interest in Figure 9.2. From the first three confidence intervals it is concluded that the effect of monocalcium phosphate in comparison to the control significantly varies between the three soil types. Furthermore, it is infered that the monocalcium phosphate effect is largest at soil type S1: the effect is at least 0.2 g/plot dry matter higher (lower bound of the first confidence interval) compared to soil type S2 and at least 1.9 g/plot dry matter higher (lower bound of the second confidence interval) compared to soil type S3. Based on these confidence intervals the applied scientist is now free to decide if these statistical significant interaction terms are also of biological relevance.

9.3. Multi-centre clinical trial

The significant treatment-by-centre interaction in Table 2.3 indicates that the treatment effect δ_i is significantly different between the five centres. The goal is now to specify if this heterogeneity of the treatment effect depends on a qualit-

ative treatment-by-centre interaction and on which centre this possible qualitative interaction is based on. Therefore the Gail and Simon test and the method proposed by Kitsche and Hothorn [2013] are applied to these data set.

Applying the Gail and Simon test to test $H_0 : \boldsymbol{\delta} \in \boldsymbol{O}^+ \cup \boldsymbol{O}^-$ results in the test statistic $Q = \min(57.75, 7.65) = 7.65$. The corresponding p-value is 0.023, and therefore, the null hypothesis that the sign of the treatment effect is equal across all of the centres can be rejected. Nevertheless, the Gail and Simon test does not admit any statement regarding the source and the amount of the significant qualitative interaction.

For a serious comparison of the proposed method with the Gail and Simon test, we conduct the analysis according to Section 4.5.1, which assumes heterogeneous variances. Nonetheless, it is recommended to make a decision regarding the assumption of variance homogeneity before the experiment is conducted.

The first row vector of $\boldsymbol{C}_{\text{Interaction}}^{\text{Numerator}}$ from Equation 3.13 and $\boldsymbol{C}_{\text{Interaction}}^{\text{Denominator}}$ from Equation 3.14 build the ratio of the treatment effect of the first centre through the overall treatment effect. Therefore the parameter γ_1 can be interpreted as the relative change of the treatment effect of the first centre to the overall treatment effect.

Figure 9.3 displays the one-sided simultaneous confidence intervals and multiplicity adjusted p-values under the assumption of heteroscedasticity to assess the null hypothesis of no qualitative interaction from Equation 5.4. According to Figure 9.3, a significant qualitative interaction in centre 101 is detected. The interpretation of the point estimators is straightforward: $\hat{\gamma}_2 = -0.58$ means that the treatment effect at centre 101 is in the opposite direction from the overall treatment effect and its amount encompasses 58% of the overall treatment effect. The treatment effect in centre 100 ($\hat{\gamma}_1 = 1.77$) is 1.77 times greater than the overall treatment effect but is not different in its direction with respect to the overall treatment effect, meaning that there is not a qualitative interaction present. Obviously, we get the same conclusion on the presence of a qualitative interaction from the Gail and Simon test and the method proposed by Kitsche and Hothorn [2013]. In addition, using the SCIs for the ratios of treatment differences allows the identification and quantification of this qualitative interaction.

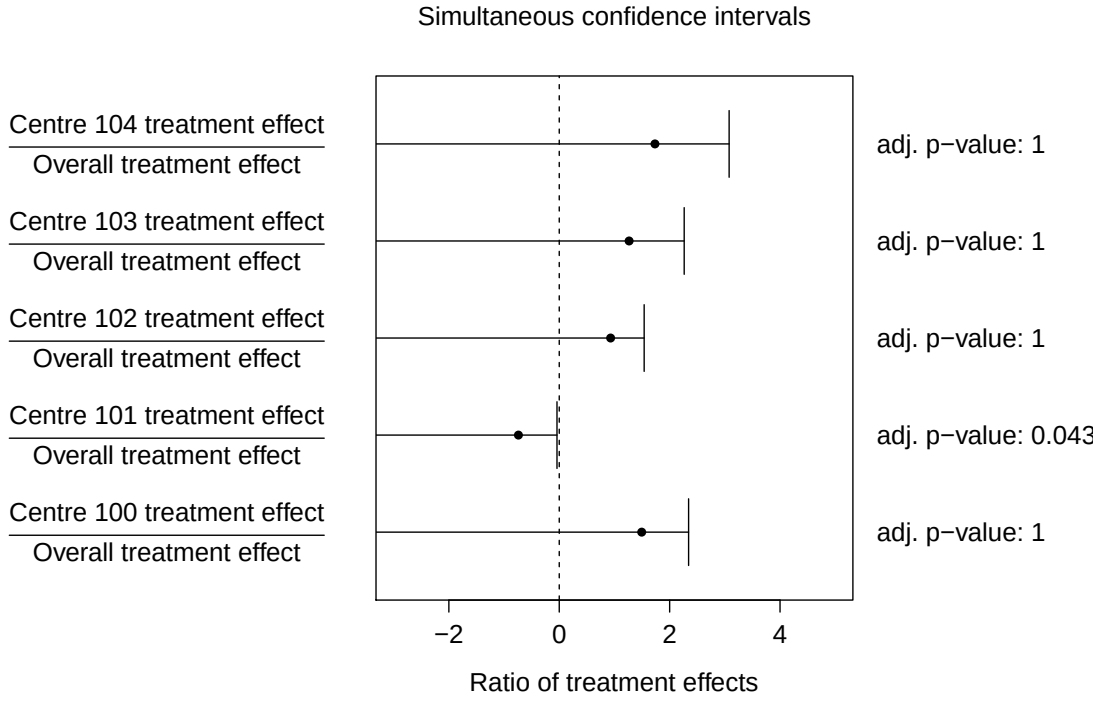

Figure 9.3.: One-sided simultaneous 95% confidence intervals for the ratios of treatment effects to assess the null hypothesis of no qualitative interaction from Equation 5.4. Points denote the point estimators $\hat{\gamma}_m$ for $m = 1, \ldots, 5$. The dashed line represents the margin defined under H_0. The upper confidence limit is displayed by a vertical line for each centre. Adjusted p-values are listed on the right-hand side of each confidence interval.

9.4. MERIT-HF study

As stated in Section 2.5, for the multi-regional MERIT-HF trial interest is in analysing the heterogeneity of the regional treatment effects in comparison to the overall treatment effect in terms of detecting a potential qualitative interaction. Therefore the Gail and Simon test and the methodology presented in Section 7 to detect a qualitative interaction for binomial data is applied to the multi-regional MERIT-HF trial.

The test statistic for the Gail and Simon test can be calculated as $Q^+ = 7.745 + 2.715 + 0.143 + 3.483 + 4.180 + 1.234 + 0.000 + 4.464 + 4.216 + 2.337 = 30.517$ and $Q^- = 0.024 + 0.078 = 0.102$. The minimum of both statistics, $\min(Q^+, Q^-)$=0.102, is smaller than the critical value 12.60 ($\alpha = 0.05, J = 12$) and therefore the null hypothesis of no qualitative interaction cannot be rejected. The corresponding p-value is 0.996.

Applying the method proposed in Section 7 to detect the source of a potential qualitative interaction the interest is now on the ratios of treatment effects:

$$\gamma_m = \frac{\boldsymbol{h}_m \boldsymbol{\pi}}{\boldsymbol{d}_m \boldsymbol{\pi}}, \qquad m = 1, \ldots, 12.$$

To determine the deviation of each region from the overall effect, the parameters γ_m are defined as the ratios of risk difference of each region to the overall risk difference. Therefore, the numerator and denominator contrast matrices are defined by using the contrast matrices defined in Equation 3.8 and 3.9. The corresponding estimated parameters γ_m, the test statistics, the multiplicity-adjusted p-values and the simultaneous upper confidence limits are presented in Table 9.1. Although the parameters $\hat{\gamma}_{Iceland} = -0.405$ and $\hat{\gamma}_{USA} = -0.140$ would suggest a qualitative interaction, we cannot reject the null hypothesis of no qualitative interaction from either the adjusted p-values or the simultaneous confidence intervals.

The observed reversal treatment effect in the US population of the MERIT-HF trial was already part of a serious discussion in the scientific literature, see, e.g., Wedel et al. [2001], Moyé [2003] and Wittes [2013]. As noted by Wedel et al. [2001], the Food and Drug Administration (FDA) decided to perform a treatment-by-country interaction. The FDA interpreted the result as in this quote from Moyé [2003]:“ *The finding of adverse United States mortality could of course be*

Table 9.1.: Estimated parameters of interest $\hat{\gamma}_m$, resulting test statistic, multiplicity-adjusted p-values to test for qualitative interaction and simultaneous upper confidence limits for the parameters γ_m in the multi-regional MERIT-HF trial.

Country	$\hat{\gamma}_m$	Test statistic	adj. p-value	Upper
Belgium	4.315	2.783	1.000	13.443
Czech Republic	1.805	1.648	1.000	6.199
Denmark/Finland	0.309	0.378	1.000	2.758
Germany	1.415	1.866	1.000	4.206
Hungary	1.721	2.045	1.000	5.082
Iceland	-0.405	-0.154	0.999	7.862
Norway	1.211	1.111	1.000	5.253
Poland	0.000	0.000	1.000	3.121
Sweden	4.076	2.113	1.000	14.064
The Netherland/Switzerland	1.200	2.053	1.000	3.577
UK	1.763	1.529	1.000	6.674
USA	-0.140	-0.279	0.997	0.984

attributable to chance, but it could alternatively be a genuine finding, the result of US-differences in demographics or concomitant therapy". The FDA handled the discordant finding by approving the drug and therefore gets the same conclusions as from the results in Table 9.1.

9.5. Trastuzumab data set

Within this subsection the dataset from the interim analysis of the trastuzumab multi-regional clinical trial presented in Section 2.5 is analysed. The goal is to assess consistency of the treatment effect, defined as the risk differences, for each region. To provide a local statement on consistency the union-intersection principle introduced in Section 6 is applied. To define appropriate hypotheses for the assessment of consistency a consistency margin ω has to be defined. The Ministry of Health, Labour and Welfare of Japan [2007] proposed that the observed treatment effect for Japanese patients should be at least half of that observed for all patients to accept consistency of the treatment. This statement can be translated into the ratio of treatment effects by defining a relative margin of $\omega = 0.5$. Then,

the local null hypothesis for the Japanese region is $H_0^{\text{Japan}} : \gamma_{\text{Japan}} \leq 0.5$. Nevertheless, as noted by Chen et al. [2010], a value of $\omega = 0.5$ may be too conservative and even not practical if more than two regions are included in the analysis. Therefore, they recommend a smaller value of $\omega = 1/J$, where J denotes the number of pre-defined regions. Within, here we adopt their approach for the definition of an appropriate consistency margin. The parameter of interest γ_m are defined as the ratio of each regional treatment effect to the overall treatment effect. Therefore the following numerator and denominator interaction contrast matrices are used.

$$\boldsymbol{C}_{\text{Interaction}}^{\text{Numerator}} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$\boldsymbol{C}_{\text{Interaction}}^{\text{Denominator}} = \begin{pmatrix} 0.05 & -0.05 & 0.12 & -0.12 & 0.10 & -0.10 & 0.03 & -0.03 & 0.70 & -0.70 \\ 0.05 & -0.05 & 0.12 & -0.12 & 0.10 & -0.10 & 0.03 & -0.03 & 0.70 & -0.70 \\ 0.05 & -0.05 & 0.12 & -0.12 & 0.10 & -0.10 & 0.03 & -0.03 & 0.70 & -0.70 \\ 0.05 & -0.05 & 0.12 & -0.12 & 0.10 & -0.10 & 0.03 & -0.03 & 0.70 & -0.70 \\ 0.05 & -0.05 & 0.12 & -0.12 & 0.10 & -0.10 & 0.03 & -0.03 & 0.70 & -0.70 \end{pmatrix}$$

The multiplicity adjusted p-values and lower confidence limits of the one-sided simultaneous confidence intervals are displayed in Figure 9.4. From the results in Figure 9.4 it can be concluded that the treatment effect in the regions Eastern Europe and Others is consistent to the overall treatment effect using a consistency margin of $\omega = 0.2$. For the Japanese region no consistency of the treatment effect can be inferred in the interim analysis. This result was also observed by Ando and Hamasaki [2010]. Nevertheless, they recommend that the evaluation should not be based on the results from the interim analysis, because the data from the two year treatment of trastuzumab were not available in the interim analysis report.

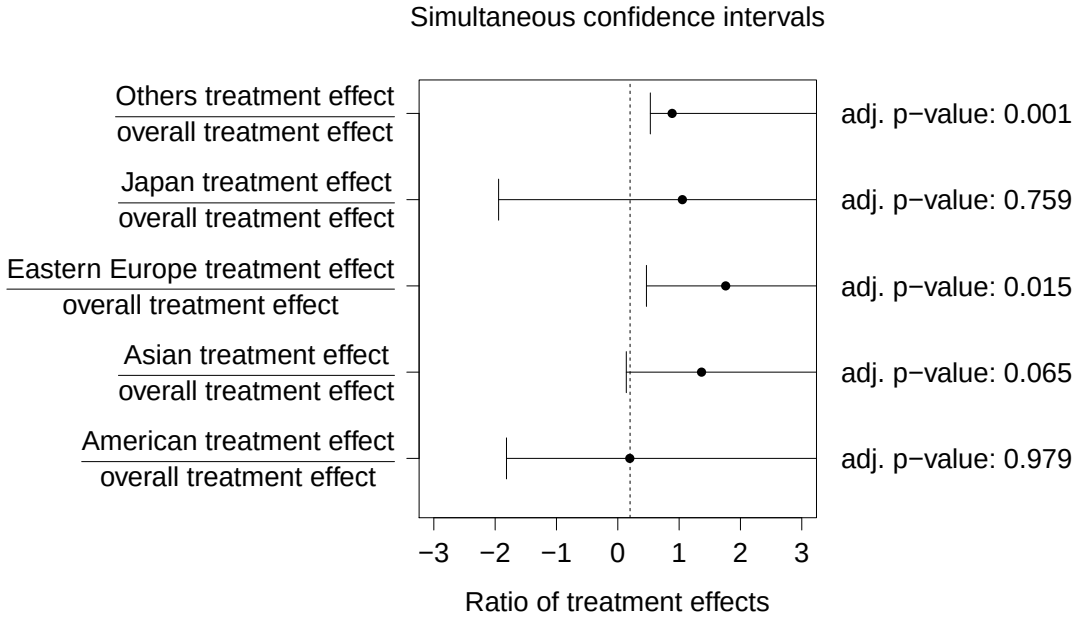

Figure 9.4.: One-sided simultaneous 95% confidence intervals for the ratios of treatment effects. Points denote the point estimators $\hat{\gamma}_m$ for $m = 1, \ldots, 5$. The dashed line represents the consistency margin $\omega = 0.2$. The lower confidence limit is displayed by a vertical line for each region. Adjusted p-values are listed on the right-hand side of each confidence interval.

Chapter 10.

Discussion

In this research, the evaluation of statistical interactions in two-factorial designs is considered with a special focus on the detection of qualitative interactions. In the first part appropriate hypotheses for an in depth analysis of interactions via product-type interaction contrasts is developed. The proposed approach allows the formulation of user specified comparisons of means that are of main interest in the experiment under consideration. The application of user-defined interaction contrasts of cell means reduces the number of tests and increases the interpretability of the results. The elaboration of the user defined interaction contrast matrices is demonstrated on five real data sets from biomedical and horticultural science. To make inferences on the developed hypotheses, multiple contrast tests that result in multiplicity adjusted p-values are considered. In addition to multiplicity adjusted p-values, the author recommend the calculation of simultaneous confidence intervals for the interaction effects. By using those confidence intervals it is possible to evaluate the direction and the magnitude of the interaction effects. This provides besides a statement on the statistical significance also a statement on the biological relevance of the interaction effects.

Furthermore, the interaction parameters are defined as the ratios of treatment effects. Depending on the formulation of the null hypothesis the ratio of treatment differences is applicable to (i) detect qualitative interactions (ii) asses the consistency of the treatment effect in a non-inferiority framework, or (iii) test for no qualitative interaction. In addition, using the ratio of treatment effects can be biologically interpreted as a percentage change of the treatment effects. It should not be unmentioned, that this method has its limitations in cases where

the denominator of the ratio of treatment effects is not significantly different from zero because the simultaneous confidence intervals are not calculable is this case. When comparing to the grand mean this case occurs when there is no main effect of the primary factor. Apart from that, using the grand mean contrast as the denominator is very powerful to detect a qualitative interaction when the sign of the treatment effect differs only in a few subsets of the secondary factor and if the overall treatment effect is small. However, the performance of the proposed method gets worse in cases where the number of subgroups can approximately be split into two groups of differing signs of the treatment effect. Nevertheless, this extreme case should be exceptional in practical trials. Alternatively, all pairwise comparisons of treatment effects can be conducted. The usage of adjusted p-values is limited for the assessment of qualitative interactions, since the test statistic reduces to a test for the numerator. Therefore, the test is only applicable with an a priori assumption on the direction of the treatment effects. This hypothesis corresponds to the one-sided Gail and Simon test.

Within this thesis, it is assumed that the primary response variable is normally distributed. The approach is further extended to the case of binary response variables. Binary response variables are very common, especially in biomedical applications, see, e.g. the illustrative examples presented within this thesis. The methods presented here assume that the binomial proportions are asymptotically normally distributed, i.e. cases of moderate to high sample sizes. Unfortunately, the situation in which the construction of simultaneous confidence intervals fails more frequently appears for the ratios of risk difference since the denominator is not significantly different from zero for small risk differences.

Further on, the presence of response variables that do not fulfil these distributional assumptions are common in applied biosciences, e.g., score data, ordinal data or continuous skewed data. In those cases, non-parametric statistical procedures can be used, that make no assumption on the underlying distribution of the data. Konietschke [2009] presented a non-parametric procedure that formulates the hypotheses of interest via linear combinations of relative effects in a one-way layout. The extension of this approach to test for statistical interaction with focus on qualitative interaction might be of interest in future research.

Finally, some remarks are given on the potential biological interpretation of a

significant qualitative interaction. In recent times, this issue is under discussion especially in the context of multi-regional trials. Pocock et al. [2013] gave some explanation for geographic inconsistencies in treatment effect in multi-regional trials, like "*type of patients recruited, their therapeutic management, and the evaluation of their outcomes*". They considered for example the Platelet Inhibition and Patient Outcomes (PLATO) trial, where the regional interaction was caused by the maintenance of aspirin [Mahaffey et al., 2011].

Bibliography

R. Abelson and D. Prentice. Contrast Tests of Interaction Hypotheses. *Psychological Methods*, 2(4):315–328, 1997.

A. Agresti. *Categorical data analysis*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken NJ, 3 edition, 2013. ISBN 0-470-46363-5.

A. Agresti and J. Hartzel. Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine*, 19(8):1115–1139, 2000.

Y. Ando and T. Hamasaki. Practical issues and lessons learned from multi-regional clinical trials via case examples: A Japanese perspective. *Pharmaceutical Statistics*, 9(3):190–200, 2010.

A. Azzalini and D. R. Cox. Two New Tests Associated with Analysis of Variance. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):335–343, 1984.

R. J. Baker. Tests for Crossover Genotype-Environmental Interactions. *Canadian Journal of Plant Science*, 68(2):405–410, 1988.

R. J. Boik. The Analysis of Two-Factor Interactions in Fixed Effects Linear Models. *Journal of Educational Statistics*, 18(1):1–40, 1993.

D. Bradu and K. R. Gabriel. Simultaneous Statistical Inference on Interactions in Two-Way Analysis of Variance. *Journal of the American Statistical Association*, 69(346):428–436, 1974.

H. Braun. *The Collected Works of John W. Tukey: Multiple Comparions*. Chapman & Hall, 1994. ISBN 9780412051210.

N. Breslow and N. Day. Statistical Methods in Cancer Research. *International Agency for Research on Cancer*, 1994.

F. Bretz. *Powerful Modifications of Williams' Test on Trend*. PhD thesis, Gottfried Wilhelm Leibniz Universitaet Hannover, 1999.

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

J. Chen, H. Quan, B. Binkowitz, S. P. Ouyang, Y. Tanaka, G. Li, S. Menjoge, and E. Ibia. Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. *Pharmaceutical Statistics*, 9(3):242–253, 2010.

J. Ciminera, J. Heyse, H. Nguyen, and J. Tukey. Tests for qualitative treatment-by-centre interaction using a Pushabck procedure. *Statistics in Medicine*, 12: 1033–1045, 1993.

M. Compton. Interaction between explant size and cultivar affects shoot organogenic competence of watermelon cotyledons. *HortScience*, 35(4):749–750, 2000.

D. R. Cox. Interaction. *International Statistical Review*, 52(1):1–31, 1984.

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

G. Dilba, F. Bretz, and V. Guiard. Simultaneous confidence sets and confidence intervals for multiple ratios. *Journal of Statistical Planning anf Inference*, 136 (8):2640–2658, 2006a.

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

G. Dilba, F. Schaarschmidt, and Hothorn L. A. Inferences for Ratios of Normal Means. *R News*, 7(1):20–23, 2007.

G. Dilba, M. Hasler, D. Gerhard, and F. Schaarschmidt. mratios: Inferences for ratios of coefficients in the general linear model, 2012. URL `http://CRAN.R-project.org/package=mratios`.

G. Djira. *Simultaneous Inference for Ratios of Location Parameters*. PhD thesis, Gottfried Wilhelm Leibniz Universitaet Hannover, 2005.

G. Djira and L. Hothorn. Detecting Relative Changes in Multiple Comparisons with an Overall Mean. *Journal of Quality Technology*, 41(1):60–65, 2009.

G. Djira, F. Schaarschmidt, and B. Fayissa. Inferences for selected location quotients with applications to health outcomes. *Geographical Analysis*, 42(3):288–300, 2010.

A. Dmitrienko and R. D'Agostino. Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, 32(29):5172–5218, 2013.

A. Dmitrienko, G. Molenberghs, C. Chuang-Stein, and W. Offen. *Analysis of Clinical Trials using SAS: A Practical Guide*. Cary, NC, SAS Institute Inc, 2005.

C. W. Dunnett. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, 50(272): 1096–1121, 1955.

E. C. Fieller. Some Problems in Interval Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):175–185, 1954.

K. R. Gabriel, J. Putter, and Y. Wax. Simultaneous Confidence Intervals for Product-Type Interaction Contrasts. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, 35(2):234–244, 1973.

M. Gail and R. Simon. Testing for Qualitative Interactions Between Treatment Effects and Patient Subsets. *Biometrics*, 41(2):361–372, 1985.

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

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2.

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. mvtnorm: Multivariate Normal and t Distributions, 2013. URL `http://CRAN.R-project.org/package=mvtnorm`.

A. B. d. Gonzalez and D. R. Cox. Interpreatation of Interactions: A Review. *The Annals of Applied Statistics*, 1(2):371–385, 2007.

S. Han, P. Rosenberg, and N. Chatterjee. Testing for Gene-Environment and Gene-Gene Interactions Under Monotonicity Constraints. *Journal of the American Statistical Association*, 107(500):1441–1452, 2012.

M. Hasler. *Extensions of Multiple Contrast Tests*. PhD thesis, Gottfried Wilhelm Leibniz Universitaet Hannover, 2009.

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

D. Hauschke and M. Kieser. Multiple Testing to Establish Noninferiority of k treatments with a Reference Based on the Ratio of Two Means. *Drug Information Journal*, 35(4):1247–1251, 2001.

J. Hirschberg and J. Lye. Two geometric representations of confidence intervals for ratios of linear combinations of regression parameters: An application to the NAIRU. *Economics Letters*, 108(1):73–76, 2010a.

J. Hirschberg and J. Lye. A Geometric Comparison of the Delta and Fieller Confidence Intervals. *The American Statistician*, 64(3):234–241, 2010b.

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

ICH E9. ICH Harmonised Tripartite Guideline, Statistical principles for Clinical Trials E9, 1998.

R. E. Kirk. *Experimental design: Procedures for the behavioral sciences*. Brooks/Cole, Pacific Grove and Calif, 3 edition, 1995. ISBN 9780534250928.

A. Kitsche and L. A. Hothorn. Testing for qualitative interaction using ratios of treatment differences. *Statistics in Medicine*, (accepted for publication), 2013. URL doi:10.1002/sim.6048.

F. Konietschke. *Simultane Konfidenzintervalle für nichtparametrische relative Kontrasteffekte*. PhD thesis, Georg-August-Universität Göttingen, 2009.

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

J. Li and I. S. F. Chan. Detecting Qualitative Interactions in Clinical Trials: An Extension of Range Test. *Journal of Biopharmaceutical Statistics*, 16(6): 831–841, 2006.

K. Mahaffey, D. Wojdyla, K. Carroll, R. Becker, R. Storey, D. Angiolillo, C. Held, C. Cannon, S. James, K. Pieper, J. Horrow, R. Harrington, and L. Wallentinon. Ticagrelor compared with clopidogrel by geographic region in the Platelet Inhibition and Patient Outcomes (PLATO) Trial. *Circulation*, 124(5):544–554, 2011.

MERIT-HF Study Group. Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trial in Congestive Heart Failure (MERIT-HF). *The Lancet*, 353(9169):2001–2007, 1999.

S. Michiels, R. F. Potthoff, and S. L. George. Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint. *Statistics in Medicine*, 30(13):1502–1518, 2011.

Ministry of Health, Labour Welfare of Japan. Basic concepts for joint international clinical trials, 2007. URL http://www.pmda.go.jp/operations/notice/2007/file/0928010-e.pdf.

L. A. Moyé. *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*. Statistics for Biology and Health. Springer, New York NY u.a, 2003. ISBN 0-387-00727-X.

H. Mukerjee, T. Robertson, and F. T. Wright. Comparison of Several Treatments with a Control Using Multiple Contrasts. *Journal of the American Statistical Association*, 82(399):p 902–910, 1987.

G. Pan and D. Wolfe. Test for qualitative interaction of clinical significance. *Statistics in Medicine*, 16:1645–1652, 1997.

R. A. Parker. Testing for qualitative interactions between stages in an adaptive study. *Statistics in Medicine*, 29(2):210–218, 2010.

R. G. Petersen. *Design and analysis of experiments*, volume 66 of *Statistics*. Dekker, New York NY u.a, 1985. ISBN 0-8247-7340-3.

R. Peto. *Treatment of Cancer*. Chapman & Hall: London, 1982. ISBN 0-340-91221-9.

S. Piantadosi and M. Gail. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine*, 12:1239–1248, 1993.

S. Pocock, G. Calvo, J. Marrugat, K. Prasad, L. Tavazzi, L. Wallentin, F. Zannad, and A. Alonso Garcia. International differences in treatment effect: Do they really exist and why? *European Heart Journal*, 34(24):1846–1852, 2013.

R. F. Potthoff, B. L. Peterson, and S. L. George. Detecting treatment-by-centre interaction in multi-centre clinical trials. *Statistics in Medicine*, 20(2):193–213, 2001.

H. Quan, M. Li, W. Shih, S. Ouyang, J. Chen, J. Zhang, and P.-L. Zhao. Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. *Statistics in Medicine*, 2012.

R Core Team. R: A Language and Environment for Statistical Computing, 2013. URL `http://www.R-project.org/`.

E. Romond, E. Perez, J. Bryant, V. Suman, C. Geyer Jr., N. Davidson, E. Tan-Chiu, S. Martino, S. Paik, P. Kaufman, S. Swain, T. Pisansky, L. Fehrenbacher, L. Kutteh, V. Vogel, D. Visscher, G. Yothers, R. Jenkins, A. Brown, S. Dakhil,

E. Mamounas, W. Lingle, P. Klein, J. Ingle, and N. Wolmark. Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine*, 353(16):1673–1684, 2005.

O. Sahin, M. Taskin, Y. Kadioglu, A. Inal, A. Gunes, and D. Pilbeam. Influence of chloride and bromate interaction on oxidative stress in carrot plants. *Scientia Horticulturae*, 137:81–86, 2012.

H. Scheffe. *The Analysis of Variance*. Wiley classics library. John Wiley & Sons, 1999. ISBN 0-471-34505-9.

S. R. Searle. *Linear models*. Wiley classics library. John Wiley & Sons, 1997. ISBN 0-471-18499-3.

K. Slauenwhite and M. Qaderi. Single and interactive effects of temperature and light quality on four canola cultivars. *Journal of Agronomy and Crop Science*, 199(4):286–298, 2013.

B. Truberg and M. Hühn. Contributions to the Analysis of Genotype x Environment Interactions: Comparison of Different Parametric and Non-parametric Tests for Interactions with Emphasis on Crossover Interactions. *Journal of Agronomy and Crop Science*, 185(4):267–274, 2000.

U.S. Department of Health, Human Services Food, and Drug Administration. ICH International Conference on Harmonization Tripartite Guidance E5 Questions and Answers: Ethnic Factor in the Acceptability of Foreign Clinical Data., 2009.

X. Wang, R. C. Elston, and X. Zhu. The Meaning of Interaction. *Human Heredity*, 70(4):269–277, 2010.

H. Wedel, D. Demets, P. Deedwania, B. Fagerberg, S. Goldstein, S. Gottlieb, A. Hjalmarson, J. Kjekshus, F. Waagstein, J. Wikstrand, and MERIT-HF Study Group. Challenges of subgroup analyses in multinational clinical trials: Experiences from the MERIT-HF trial. *American Heart Journal*, 142(3):502–511, 2001.

S. Wellek. Testing for absence of qualitative interactions between risk factors and treatment effects. *Biometrical Journal*, 39(7):809–821, 1997.

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

J. Wittes. *Why Is This Subgroup Different from All Other Subgroups? Thoughts on Regional Differences in Randomized Clinical Trials, Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*. Lecture Notes in Statistics. Springer, 2013.

Appendix A.

R Code for reproducible research

A.1. Bush beans data set

```
#generating the data set and assign it to the object "Beans"
Beans <- data.frame(
  Variety =  c("NewEra",    "NewEra",    "NewEra",    "BigGreen", "BigGreen", "BigGreen",
               "LittleGem", "LittleGem", "LittleGem", "RedLake",  "RedLake",  "RedLake",
               "NewEra",    "NewEra",    "NewEra",    "BigGreen", "BigGreen", "BigGreen",
               "LittleGem", "LittleGem", "LittleGem", "RedLake",  "RedLake",  "RedLake",
               "NewEra",    "NewEra",    "NewEra",    "BigGreen", "BigGreen", "BigGreen",
               "LittleGem", "LittleGem", "LittleGem", "RedLake",  "RedLake",  "RedLake",
               "NewEra",    "NewEra",    "NewEra",    "BigGreen", "BigGreen", "BigGreen",
               "LittleGem", "LittleGem", "LittleGem", "RedLake",  "RedLake",  "RedLake"),
  Spacing =  c("20", "40", "60", "20", "40", "60", "20", "40", "60", "20", "40", "60",
               "20", "40", "60", "20", "40", "60", "20", "40", "60", "20", "40", "60",
               "20", "40", "60", "20", "40", "60", "20", "40", "60", "20", "40", "60",
               "20", "40", "60", "20", "40", "60", "20", "40", "60", "20", "40", "60"),
  Block =   c("I",   "I",   "I",   "I",   "I",   "I",   "I",   "I",   "I",   "I",   "I",   "I",
              "II",  "II",  "II",  "II",  "II",  "II",  "II",  "II",  "II",  "II",  "II",  "II",
              "III", "III", "III", "III", "III", "III", "III", "III", "III", "III", "III", "III",
              "IV",  "IV",  "IV",  "IV",  "IV",  "IV",  "IV",  "IV",  "IV",  "IV",  "IV",  "IV"),
  Yield =   c(32, 36, 42, 37, 39, 50, 35, 34, 33, 40, 35, 28, 21, 26, 33, 38,
              45, 54, 32, 33, 29, 36, 33, 28, 19, 21, 26, 27, 44, 54, 29, 28,
              25, 35, 31, 23, 22, 24, 26, 30, 37, 42, 30, 28, 26, 38, 35, 30))
str(Beans)#display the structure of the data set
Beans$Variety <-factor(Beans$Variety, levels=c("BigGreen", "NewEra", "LittleGem", "RedLake"))
#reorder the levels of the factor Variety
#generate a new factor variable that combines the variable Variety and Space
#to get a pseudo-one-way layout
Beans$VarSpace <- factor(Beans$Variety : Beans$Spacing)
#fitting a linear model and calculate an ANOVA
fm <- lm(Yield ~ Variety * Spacing + Block, data = Beans)
anova(fm)
```

```
#fitting a cell means model
CellMeansModel <- lm(Yield ~ VarSpace + Block -1, data = Beans)
#define appropriate user defined contrast matrices
VarMat <- matrix(c(0.5,  0.5, -0.5, -0.5,
                   1,   -1,    0,    0,
                   0,    0,    1,   -1), nrow=3, byrow=TRUE)
SpaceMat <- matrix(c(1, -1,  0,
                   1,  0, -1,
                   0,  1, -1), nrow=3, byrow=TRUE)
#define the interaction contrast matrix by building the Kronecker product
#of the previously defined contrast matrices
InteractionMat <- kronecker(VarMat, SpaceMat)
library(multcomp)#add on package multcomp is required for multiple comparisons
MultTest <- glht(model=CellMeansModel, linfct = mcp(VarSpace=InteractionMat))
summary(MultTest)#calculating adjusted p-values
confint(MultTest)#calculating simultaneous confidence intervals
```

A.2. Lettuce data set

```
#generating the data set and assign it to the object "Lettuce"
Lettuce <- data.frame(
  Soil =       c("S1", "S1", "S1", "S1", "S2", "S2", "S2", "S2", "S3", "S3", "S3", "S3",
                 "S1", "S1", "S1", "S1", "S2", "S2", "S2", "S2", "S3", "S3", "S3", "S3",
                 "S1", "S1", "S1", "S1", "S2", "S2", "S2", "S2", "S3", "S3", "S3", "S3",
                 "S1", "S1", "S1", "S1", "S2", "S2", "S2", "S2", "S3", "S3", "S3", "S3"),
  Fertilizer = c( "control"      ,"control"      ,"control"      ,"control"      ,"control"
                 ,"control"      ,"control"      ,"control"      ,"control"      ,"control"
                 ,"control"      ,"control"      ,"cal.phos"     ,"cal.phos"     ,"cal.phos"
                 ,"cal.phos"     ,"cal.phos"     ,"cal.phos"     ,"cal.phos"     ,"cal.phos"
                 ,"cal.phos"     ,"cal.phos"     ,"cal.phos"     ,"cal.phos"     ,"pot.metaphos"
                 ,"pot.metaphos" ,"pot.metaphos" ,"pot.metaphos" ,"pot.metaphos" ,"pot.metaphos"
                 ,"pot.metaphos" ,"pot.metaphos" ,"pot.metaphos" ,"pot.metaphos" ,"pot.metaphos"
                 ,"pot.metaphos" ,"magn.phos"    ,"magn.phos"    ,"magn.phos"    ,"magn.phos"
                 ,"magn.phos"    ,"magn.phos"    ,"magn.phos"    ,"magn.phos"    ,"magn.phos"
                 ,"magn.phos"    ,"magn.phos"    ,"magn.phos"),
  Weight =     c(0.49, 0.31, 0.21, 1.10, 0.54, 0.20, 0.13, 1.01, 2.71, 2.00, 1.95, 2.62, 3.89,
                 3.78, 3.03, 3.81, 2.47, 2.30, 1.64, 2.50, 2.22, 2.40, 1.93, 2.88, 3.52, 3.80,
                 2.96, 3.81, 2.18, 1.90, 1.49, 2.42, 3.01, 3.14, 2.78, 3.72, 3.67, 4.41, 3.76,
                 4.40, 1.66, 1.82, 1.64, 2.49, 2.19, 2.69, 2.08, 2.92))
str(Lettuce)#display the structure of the data set
Lettuce$Fertilizer <-factor(Lettuce$Fertilizer,
                            levels=c("control", "cal.phos", "pot.metaphos", "magn.phos"))
#reorder the levels of the factor Fertilizer
#generate a new factor variable that combines the variable Fertilizer and Soil
#to get a pseudo-one-way layout
Lettuce$FertSoil <- factor(Lettuce$Fertilizer:Lettuce$Soil)
```

```
#fitting a linear model and calculate an ANOVA
fm <- lm(Weight ~ Fertilizer * Soil, data = Lettuce)
anova(fm)
#fitting a cell means model
CellMeansModel <- lm(Weight ~ FertSoil -1, data = Lettuce)
#define appropriate user defined contrast matrices
SoilMatrix <- matrix(c(1, -1, 0,
                       1,  0,-1,
                       0,  1,-1), nrow=3, byrow=TRUE)
FertMatrix <- matrix(c(1, -1, 0, 0,
                       1,  0,-1, 0,
                       1,  0, 0,-1),nrow=3, byrow=TRUE)
#define the interaction contrast matrix by building the Kronecker produkt of the
#previously defined contrast matrices
InteractionMat <- kronecker(FertMatrix,SoilMatrix)
rownames(InteractionMat) <- c("(Cont. - Mono.)S1 - (Cont. - Mono.)S2",
                              "(Cont. - Mono.)S1 - (Cont. - Mono.)S3",
                              "(Cont. - Mono.)S2 - (Cont. - Mono.)S3",
                              "(Cont. - Magn.)S1 - (Cont. - Magn.)S2",
                              "(Cont. - Magn.)S1 - (Cont. - Magn.)S3",
                              "(Cont. - Magn.)S3 - (Cont. - Magn.)S3",
                              "(Cont. - Pott.)S1 - (Cont. - Pott.)S2",
                              "(Cont. - Pott.)S1 - (Cont. - Pott.)S3",
                              "(Cont. - Pott.)S2 - (Cont. - Pott.)S3")
#add on package multcomp is required for multiple comparisons
library(multcomp)
MultTest <- glht(model=CellMeansModel, linfct = mcp(FertSoil=InteractionMat))
summary(MultTest)#calculating adjusted p-values
confint(MultTest)#calculating simultaneous confidence intervals
```

A.3. Multi-centre clinical trial

```
#Data set
Depression <- data.frame(
    Centre =    c(100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100,
                  100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 101, 101,
                  101, 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, 102,
                  102, 102, 102, 102, 102, 102, 102, 102, 102, 102, 102, 102,
                  102, 102, 102, 102, 102, 102, 102, 102, 102, 102, 102, 102,
                  102, 102, 102, 103, 103, 103, 103, 103, 103, 103, 103, 103,
                  103, 103, 103, 103, 103, 103, 103, 103, 103, 104, 104, 104, 104,
                  104, 104, 104, 104, 104, 104, 104, 104, 104),
    Group =     c("P", "P", "D", "D", "P", "P", "D", "D", "P", "P", "D", "D", "P",
                  "P", "D", "D", "P", "P", "D", "P", "P", "P", "D", "D", "P", "P",
                  "D", "D", "P", "P", "D", "D", "P", "P", "D", "D", "P", "D", "P",
                  "P", "P", "P", "D", "D", "P", "P", "D", "D", "P", "P", "D", "D",
```

```
                "P", "P", "D", "D", "P", "P", "D", "D", "P", "P", "D", "D", "D",
                "D", "D", "D", "P", "P", "D", "D", "P", "P", "D", "D", "P", "P",
                "D", "D", "P", "P", "D", "D", "P", "P", "D", "D", "P", "P", "D",
                "D", "P", "P", "D", "D", "P", "P", "D", "D"),
    Score =     c(18, 14, 23, 18, 10, 17, 18, 22, 13, 12, 28, 21, 11,  6, 11, 25,
                   7, 10, 29, 12, 12, 10, 18, 14, 18, 15, 12, 17, 17, 13, 14,  7,
                  18, 19, 11,  9, 12, 11, 18, 15, 12, 18, 20, 18, 14, 12, 23, 19,
                  11, 10, 22, 22, 19, 13, 18, 24, 13,  6, 18, 26, 11, 16, 16, 17,
                   7, 19, 23, 12, 16, 11, 11, 25,  8, 15, 28, 22, 16, 17, 23, 18,
                  11, -2, 15, 28, 19, 21, 17, 13, 12,  6, 19, 23, 11, 20, 21, 25,
                   9,  4, 25, 19))
#define variable Centre as factor:
Depression$Centre <- as.factor(Depression$Centre)
#define a new variable for group-by-centre combination:
Depression$GroupCentre <- with(Depression, Centre:Group)
#calculation of the sample size for each centre over the groups:
SampleSizeCentre <- with(Depression,aggregate(Score ~ Centre, FUN=length))[,2]
#define the contrast matrices
library(mratios)
C_Drug   <- matrix(c(1,-1), nrow=1)
C_Centre_Numerator   <- contrMatRatio(n=SampleSizeCentre, type = "GrandMean")$numC
C_Centre_Denominator <- contrMatRatio(n=SampleSizeCentre, type = "GrandMean")$denC
C_Numerator   <- kronecker(C_Centre_Numerator, C_Drug)
C_Denominator <- kronecker(C_Centre_Denominator, C_Drug)
#calculation of the ratios of treatment effects assuming heterogeneous variances
Sim_Conf_Int_Hetero <- sci.ratioVH(Score ~ GroupCentre, data= Depression,
                                   Num.Contrast=C_Numerator,
                                   Den.Contrast=C_Denominator,
                                   alternative="less")
Adj_P_Values_Hetero <- simtest.ratioVH(Score ~ GroupCentre, data= Depression,
                                       Num.Contrast=C_Numerator,
                                       Den.Contrast=C_Denominator,
                                       Margin.vec=0,
                                       alternative="less")
```

A.4. MERIT-HF study

```
#Analysis of data from Quan et al. (2012)
Country <- c("Belgium", "Czech Republic", "Denmark/Finland","Germany","Hungary",
             "Iceland","Norway","Poland","Sweden","The Netherland/Switzerlnd",
             "UK","USA")
Treatment <- c("Meto CR/XL","Placebo")
SampleSize <- c(68,66,123,124,161,164,252,247,211,212,19,22
          ,97,105,102,102,39,46,299,291,87,83,532,539)
Events <- c(3,13,9,17,11,13,19,31,16,29,2,2,6,11,8,8,2,9,14,26,4,9,51,49)
Data <- data.frame(Country=rep(Country,each=2),
```

```
                  Treatment=rep(Treatment,12),
                  SampleSize=SampleSize,
                  Events=Events)
#generating binary response data set
BinaryVec       <- as.integer()
BinaryRegion    <- factor()
BinaryTreatment <- factor()
for(i in 1:length(Data[,1])){
  BinaryVec <- c(BinaryVec, rep(1, Data[i,4]))
  BinaryVec <- c(BinaryVec, rep(0, Data[i,3]-Data[i,4]))
  BinaryRegion <- c(BinaryRegion, rep(levels(factor(Data[i,1])), Data[i,3]))
  BinaryTreatment <- c(BinaryTreatment, rep(levels(factor(Data[i,2])), Data[i,3]))
}
BinaryData <- data.frame(Region = BinaryRegion,
                         Treatment = BinaryTreatment,
                         Success = BinaryVec)
BinaryData$RegionTreat <- with(BinaryData, Region:Treatment)
#Fitting with MCPAN package
library(MCPAN)
library(mratios)
MCPAN_Est <- binomest(Success ~ RegionTreat,data=BinaryData, success="1", method="Wald")
cmat <- diag(rep(1,length(levels(BinaryData$RegionTreat))))
#Variance Covariance Matrix for proportions
VarCovMat <- diag(MCPAN_Est$varp)
#Vector of estimated proportions
EstProp <- MCPAN_Est$estp
#Definition of contrast matrices
SampleSizeRegion <- with(BinaryData, aggregate(Success ~ Region, FUN=length))[,2]
C_Treatment <- matrix(c(-1,1), nrow=1)
C_Region_Numerator <- contrMatRatio(n=SampleSizeRegion, type = "GrandMean")$numC
C_Region_Denominator <- contrMatRatio(n=SampleSizeRegion, type = "GrandMean")$denC
C_Numerator <- kronecker(C_Region_Numerator, C_Treatment)
C_Denominator <- kronecker(C_Region_Denominator, C_Treatment)
#Calculating the degrees of freedom
SampleSizes <- with(BinaryData, aggregate(Success ~ Region+Treatment, FUN=length))[,3]
DF <- sum(SampleSizes-1)
ConfIntProp <- gsci.ratio(est=EstProp,
           vcmat=VarCovMat,
           Num.Contrast=C_Numerator,
           Den.Contrast=C_Denominator,
           degfree = DF,
           conf.level = 0.95,
           alternative = "less",
           adjusted = TRUE)
ConfIntProp
```

A.5. Trastuzumab data set

```
#Analysis of data from  Y. Ando and T. Hamasak (2010)
Country <- c("Japan","Asia","EasternEurope","America","Others")
Treatment <- c("Trastuzmab","Observation")
SampleSize <- c(41,46,202,202,189,175,94,94,1208,1222)
Events <- c(3,6,12,27,10,26,7,8,98,158)
Data <- data.frame(Country=rep(Country,each=2),
                   Treatment=rep(Treatment,5),
                   SampleSize=SampleSize,
                   Events=Events)
Data
#generating binary response data set
BinaryVec       <- as.integer()
BinaryRegion    <- factor()
BinaryTreatment <- factor()
for(i in 1:length(Data[,1])){
  BinaryVec <- c(BinaryVec, rep(1, Data[i,4]))
  BinaryVec <- c(BinaryVec, rep(0, Data[i,3]-Data[i,4]))
  BinaryRegion <- c(BinaryRegion, rep(levels(factor(Data[i,1])), Data[i,3]))
  BinaryTreatment <- c(BinaryTreatment, rep(levels(factor(Data[i,2])), Data[i,3]))
}
BinaryData <- data.frame(Region = BinaryRegion,
                         Treatment = BinaryTreatment,
                         Success = BinaryVec)
BinaryData$RegionTreat <- with(BinaryData, Region:Treatment)
aggregate(Success ~ Region+Treatment ,data=BinaryData, FUN=sum)
#Fitting with MCPAN package
library(MCPAN)
library(mratios)
#Assessment of local consistency
MCPAN_Est <- binomest(Success ~ RegionTreat,data=BinaryData, success="1", method="Wald")
cmat <- diag(rep(1,length(levels(BinaryData$RegionTreat))))
Waldci(cmat=cmat,
       estp=MCPAN_Est$estp,
       varp=MCPAN_Est$varp,
       varcor=MCPAN_Est$varp)
#Variance Covariance Matrix for proportions
VarCovMat <- diag(MCPAN_Est$varp)
#Vector of estimated proportions
EstProp <- MCPAN_Est$estp
#Definition of contrast matrices
SampleSizeRegion <- with(BinaryData, aggregate(Success ~ Region, FUN=length))[,2]
C_Treatment <- matrix(c(1,-1), nrow=1)
C_Region_Numerator <- contrMatRatio(n=SampleSizeRegion, type = "GrandMean")$numC
C_Region_Denominator <- contrMatRatio(n=SampleSizeRegion, type = "GrandMean")$denC
C_Numerator <- kronecker(C_Region_Numerator, C_Treatment)
C_Denominator <- kronecker(C_Region_Denominator, C_Treatment)
#Calculating the degrees of freedom
SampleSizes <- with(BinaryData, aggregate(Success ~ Region+Treatment, FUN=length))[,3]
DF <- sum(SampleSizes-1)
```

```
ConfIntProp <- gsci.ratio(est=EstProp,
                          vcmat=VarCovMat,
                          Num.Contrast=C_Numerator,
                          Den.Contrast=C_Denominator,
                          degfree = DF,
                          conf.level = 0.95,
                          alternative = "greater",
                          adjusted = TRUE)
ConfIntProp
```

Lebenslauf

Persönliche Daten

Name	Andreas Kitsche
Geboren	10.02.1986 in Zerbst
Staatsangehörigkeit	deutsch

Wissenschaftlicher Werdegang

Sep. 2010 - Februar 2014	Gottfried Wilhelm Leibniz Universität Hannover Promotion am Institut für Biostatistik Thema der Promotion: "Evaluation of interaction effects in two-factorial designs by simultaneous confidence intervals in the cell means model"
Okt. 2008 - Aug. 2010	Gottfried Wilhelm Leibniz Universität Hannover Masterstudiengang: Pflanzenbiotechnologie Schwerpunkte: Biostatistik, Molekularbiologie Abschluss: Master of Science Thema der Masterarbeit: "Einbeziehung von historischen Daten in die statistische Auswertung von Bioassays in der Toxikologie"
Okt. 2005 - Sep. 2008	Gottfried Wilhelm Leibniz Universität Hannover Bachelorstudiengang: Pflanzenbiotechnologie Abschluss: Bachelor of Science Schwerpunkte: Biotechnologie, Molekularbiologie, Biostatistik Thema der Bachelorarbeit: "Anzucht von einzelligen Grünalgen unter optimierten Bedingungen für die Produktion von Carotinoiden"
Juni 2005	Abitur am Gymnasium Francisceum Zerbst

Hannover, den 19.02.2014 Andreas Kitsche

Publikationsliste

Rezensierte Zeitschriften

- **Kitsche A** (2014): Detecting qualitative interactions in clinical trials with binary responses *Pharmaceutical Statistics* major revision
- **Kitsche A, Schaarschmidt F** (2014): Analysis of statistical interactions in factorial experiments *Journal of Agronomy and Crop Science* accepted with minor revision
- **Jaki T, Kitsche A, Hothorn LA** (2014) Statistical evaluation of toxicological assays with zero or near-to-zero proportions or counts in the concurrent negative control group: A tutorial *JP Journal of Biostatistics* accepted for publication
- **Kitsche A, Hothorn LA** (2013): Testing for qualitative interaction using ratios of treatment differences *Statistics in Medicine* DOI: 10.1002/sim.6048 accepted for publication
- **Kitsche A, Kalesse M** (2013): Configurational Assignment of Secondary Hydroxyl Groups and Methyl Branches in Polyketide Natural Products through Bioinformatic Analysis of the Ketoreductase Domain, *ChemBioChem* 14(7):851-861
- **Kitsche A, Schaarschmidt F, Hothorn LA** (2012): The use of historical controls in estimation of simultaneous confidence intervals for comparisons against a concurrent control, *Computational Statistics and Data Analysis* 56(12):3865-3875

Konferenzbeiträge

- **Kitsche A** Testing for qualitative interaction using ratios of treatment differences *International Conference on Simultaneous Inference* Hannover, Germany, September 24-26, 2013, Vortrag
- **Kitsche A** Assessment of the Heterogeneity of the Treatment Effect among Subgroups by Detecting Qualitative Interactions *XXVIth International Biometric Conference* Kobe, Japan, August 26-31, 2012, Vortrag
- **Kitsche A** Evaluation of interaction effects with user defined contrasts in the cell means model *2nd conference of the Central European Network* Zurich, Switzerland, September 12-16, 2011, Poster
- **Kitsche A, Schaarschmidt F** Simultaneous confidence intervals incorporating historical control data *Non-Clinical Statistics Conference* Lyon, France, September 27-29, 2010, Poster

Danksagung

An dieser Stellung möchte ich mich bei denjenigen bedanken, die mich während meiner Promotionsphase begleitet haben. Mein besonderer Dank gilt meinem Doktorvater Prof. Dr. L.A. Hothorn der mir mit seinem Fachwissen zur Seite stand und mir immer wieder neue thematische Anregungen mit auf dem Weg brachte. Ebenso danke ich Prof. Dr. H.-P. Piepho für seine Bereitschaft zur gutachterlichen Stellungnahme zu dieser Arbeit. Weiterhin möchte ich mich bei Herrn Prof. Dr. M. Kalesse für seine Zusammenarbeit bei in einem interessanten Forschungsprojekt bedanken und für seine Bereitschaft den Prüfunsgvorsitz bei meiner Disputation zu übernehmen.

Bedanken möchte ich mich auch bei den Mitarbeitern des Instituts für Biostatistik für eine Vielzahl an informativen und lustigen Gesprächen. Außerdem danke ich meiner Familie für ihre stetige moralische Unterstützung.

Die Arbeit an diesem Projekt wurde mit Hilfe von Mitteln aus dem Projekt der Deutschen Froschungsgemeinschaft DfG-HO1687 "Simultane Konfidenzintervalle für nichtparametrische Effekte in faktoriellen Modellen" finanziell unterstützt.