Aus dem Institut für Medizinische Statistik,
Informatik und Epidemiologie der Universität zu Köln
Direktor: Universitätsprofessor Dr. rer. nat. W. Lehmacher

Bootstrap-Methoden für multifaktorielle Dosis-Wirkungs-Beziehungen

Inaugural-Dissertation zur Erlangung der Würde eines doctor rerum medicinalium
der Hohen Medizinischen Fakultät der Universität zu Köln

vorgelegt von Dipl.-Math. Peter Frommolt aus Nordhorn

Promoviert am 25. März 2009

Gedruckt mit Genehmigung der Medizinischen Fakultät der Universität zu Köln
2009

Druck: Hundt Druck GmbH, Köln

Dekan:	Universitätsprofessor Dr. med. J. Klosterkötter
1. Berichterstatter:	Privatdozent Dr. rer. medic. M. Hellmich
2. Berichterstatter:	Universitätsprofessor Dr. rer. nat. W. Lehmacher

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Herstellung des Manuskriptes habe ich Unterstützungsleistungen von folgenden Personen erhalten:

Privatdozent Dr. rer. medic. M. Hellmich
Universitätsprofessor Dr. rer. nat. W. Lehmacher

Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorliegenden Dissertation stehen.

Die Arbeit wurde von mir bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und ist auch noch nicht veröffentlicht.

_____

Köln, den 10.09.2008

Danksagung

Ich danke allen Kollegen und Freunden, die mich bei der Anfertigung dieser Dissertationsschrift fachlich und menschlich unterstützt haben, besonders aber

Prof. Dr. Walter Lehmacher für stetige Förderung und angeregte fachliche Diskussionen,

Priv.-Doz. Dr. Martin Hellmich für die Heranführung an das Thema, angeregte fachliche Diskussionen über den Rahmen dieser Arbeit hinaus, geduldige Vermittlung von vielerlei Kenntnissen und die Bereitstellung des R-Codes zur Auswertung der Methoden seiner Veröffentlichung von 2005,

meinen Eltern Dr. Andreas und Maria Frommolt für ständige Unterstützung jeder Art während meiner gesamten Ausbildungszeit sowie **Dr. Ruth Lohr und Anita Haapamäki** für offene Ohren und geschwisterlichen Beistand.

Diese Arbeit wurde in englischer Sprache verfasst, um den Leserkreis zu erweitern und die Publikation in einer Fachzeitschrift vorzubereiten.

Table of Contents

1 Introduction

Bifactorial trial designs are used to test for the efficacy of fixed combinations of two component drugs which form a large proportion of the total market. The efficacy mechanisms are typically different for both agents and thus additive or even synergistic effects are expected which allow application of lower doses, decreasing the risk of adverse reactions and toxicity. Clinically, it is of common interest if the respective combination groups have a significantly higher response than both mono therapy groups as otherwise the use of a drug combination would mean an unnecessary exposure to additional medication and the risk of undesired interactions or side effects. In the drug admission process, statistical confirmation of this property is required by regulatoric authorities as the European Medicines Evaluation Agency (EMEA) and the United States Food and Drug Administration (FDA).

The analysis of dose-response relationships of combination drugs is needed for dose finding purposes, i.e. to identify appropriate drug combinations in phase II trials as well as for combined phase II/III trials. As an example of such an experiment, Hung (2000) reported the results of a bifactorial clinical trial on hypertension patients who received a combination of a diuretic (drug A) and an ACE inhibitor (drug B). The primary efficacy parameter was the mean decrease in sitting diastolic blood pressure (SiDBP) with the response means and sample size allocation (in parentheses) summarized as follows:

	A=0	A=1	A=2	A=3
B=0	0 (75)	1.4 (75)	2.7 (74)	4.6 (48)
B=1	1.8 (74)	2.8 (75)	5.7 (74)	8.2 (49)
B=2	2.8 (48)	4.5 (50)	7.2 (48)	10.9 (48)

A pooled standard deviation of $\hat{\sigma} = 7.07$ was estimated.

The response means can be displayed as a function of the dose combinations in a three-dimensional plot as shown in Figure 1.1a. Statistical tests are applied to the question whether a particular combination is significantly more efficacious than both component drugs. The AVE- and MAX-tests proposed by Hung, Chi and Lipicky (1993) and Hung (2000) give an answer to the question if there exists *at least one* combination drug with this property and can be applied to the above example. Hellmich and Lehmacher (2005) reported $p_{ave} = 0.000011$ and $p_{max} = 0.000048$ for this. Now, these results both indicate that there exists at least one combination that is significantly superior to both of its

components. The question for *which* particular subset of combinations this is true leads to a multiple hypotheses problem for which the min-test proposed by Laska and Meisner (1989) applies. Hellmich and Lehmacher (2005) determined one p-value for each combination group:

	A=1	A=2	A=3
B=1	0.69	0.029	0.035
B=2	0.50	0.007	0.000048

These *p*-values have been adjusted to account for the inflation of the type I error: if multiple hypotheses are tested, the probability to reject at least one true null hypothesis is in general greater than the nominal level but can be controlled by adjustment of the corresponding *p*-values. The result from this analysis is that the combinations $(1,2)$, $(1,3)$, $(2,2)$ and $(2,3)$ are all significantly more efficacious than their respective components.

Consider as a second application the example from the paper of Huang et al. (2007) where the primary efficacy parameter was a dichotomous variable. The authors report on a parallel phase I/II trial which is currently underway and was designed to evaluate safety and efficacy for a combination of low-dose decitabine with cytarabine in the treatment of acute myeloid leukemia (AML). Therapy response is investigated as the binary endpoint of this trial, where achievement of complete remission is considered as the response criterion. Toxicity and efficacy profiles of decitabine alone have been determined in a phase I trial supervised by Issa et al. (2004) that surprisingly showed the most efficacious levels to be those for low-dose decitabine; these can be described as two levels such as “low-dose 1” and “low-dose 2”. According to the evaluation of complete remission under cytarabine alone, one “low-dose” and one “high-dose” level are studied. The response rate for the latter is taken from Petersdorf et al. (2007), whereas the analysis will be based on a guessed value for the low-level group.
As recruiting has just begun in this bifactorial trial, data for the response fraction under combination therapy of decitabine and cytarabine are not available yet. Taking the results of Issa et al. (2004) and Petersdorf et al. (2007) together, imaginary response rates are guessed according to the single-compound treatments. For the methodology exemplified here it is not relevant if the response rates which the calculations are based on exactly reflect the truth. A possible scenario for the outcome of the bifactorial trial might look similar to the following table with drug A representing decitabine and B representing cytarabine. A placebo group with a zero rate is added which is needless for the min-test analysis but required for placebo-controlled evaluation of the component drugs.

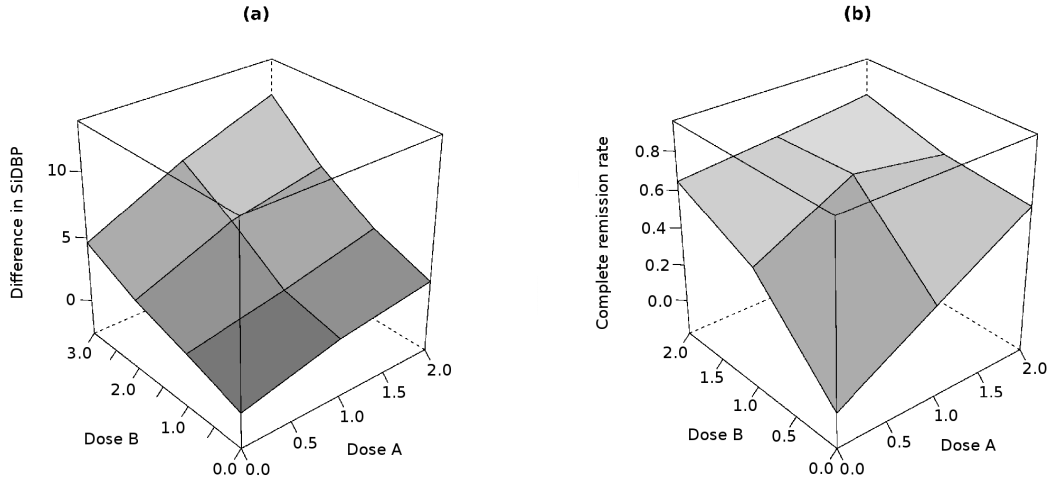

Figure 1.1: **(a)** Graphical representation of the hypertension example from Hung (2000). The reduction of sitting diastolic blood pressure (SiDBP) is displayed as the primary efficacy parameter. **(b)** Corresponding vizualisation of the binary data example from Huang et al. (2007). The proportion of individuals that achieve complete remission was determined in a sample of acute myeloid leukemia (AML) patients undergoing a combination therapy. The figures have been generated by the R package `bifactorial`.

The following table specifies the respective response fractions per group and sample size in parentheses:

	(A,0)	(A,1)	(A,2)
(B,0)	0.00 (50)	0.45 (31)	0.65 (17)
(B,1)	0.30 (100)	0.71 (50)	0.70 (50)
(B,2)	0.59 (101)	0.64 (50)	0.75 (50)

These data are displayed in Figure 1.1b similarly as for the previous example. For the binary case, there is no complete statistical methodology available up to now, at least not as detailed as for the continuous case. The analysis of this example is therefore deferred to Chapter 5 where binary endpoints are studied.

For continuous data, it was pointed out by Hung, Chi and Lipicky (1993) that the power of the statistical methods for such a design strongly depends on how well the efficacy differs between the respective two component drugs. Furthermore, the theory is restricted to standard distributional cases and requires quite technical derivations of the test statistics' distribution functions to determine the p-values. For practical purposes, it is also

desirable to complement the p-values by corresponding confidence intervals and to introduce concepts for sample size planning. In this thesis, bootstrap-based methods will be applied to these problems as they allow for arbitrary distributional properties of the data and their implementation is comparatively simple.

2 A general approach to MCP and resampling methods

Multiple comparison procedures (MCP) are important for many practical applications. Let $\mathcal{H}_0 = \{H_{0i}\}_{i \in I}$ be a family of hypotheses, where I is a finite set. Typically, tests for particular formations of parameters or some arbitrary subcollection of tests might be of interest that both can be represented as *multiple contrasts* by a vector of coefficients from the linear space $\mathbb{C}^m := \{\mathbf{c} \in \mathbb{R}^m | \sum_{i=1}^m c_i = 0\}$. Anyway, regarding the relation to the following chapters, the focus will from now on primarily be on pairwise comparisons on a single parameter between a particular subcollection of treatment groups, which is obtained as a special case of multiple contrasts. If m groups with data $\mathbf{X}_1, ..., \mathbf{X}_m$ are involved in the analysis, this results in a maximum number of $k = \binom{m}{2}$ tests. Formally, the index set I can then be written as a subset $I \subseteq \{1, ..., m\} \times \{1, ..., m\}$ such that an example of $\mathcal{H}_0$ might be a collection of test problems with $i = (i_1, i_2) \in I$ and

$$H_{0i} : \mu_{i_1} - \mu_{i_2} = 0 \quad \text{vs.} \quad H^1 : \mu_{i_1} - \mu_{i_2} \neq 0, \tag{2.1}$$

where in the treatment groups (i_1, i_2), μ_{i_1} and μ_{i_2} denote a parameter of interest in connection with the research question, e.g. means of normal or event rates in binomial populations, respectively. Furthermore, let $\mathbf{H}_0^I := \bigcap_{i \in I} H_{0i}$ and $\mathbf{H}_0^J := \bigcap_{i \in J} H_{0i}$ for $J \subset I$ be the *complete* and *partial* null hypotheses.

As an issue that is practically important and relevant to the applications discussed in this thesis, all methods are discussed for multiple t-tests but also apply to χ^2-type and other statistics in an analogous way. If tests on equality of certain population means are to be performed as in the application of Chapter 3, the decision on the system $\mathcal{H}_0$ is typically based on k two-sample t-tests with a vector $\mathbf{T} := (T_1, ..., T_k) \in \mathbb{R}^k$ of test statistics. Under the null hypothesis, each p-value in the analysis is uniformly distributed on the interval $[0, 1]$, i.e. containing p is equally likely for each of the 20 intervals $[0, .05), ..., [.95, 1]$ provided that the data are independent and normally distributed with equal variances in the respective two groups. The type I error then exactly matches the given significance level of 0.05. On the other hand, consider the event that *at least one* p-value is smaller than α though all hypotheses are true, or equivalently, the latter holds for the *minimum* of k p-values:

$$P(\min_{1 \leq i \leq k} p_i \leq 0.05) = 1 - P(p_i > 0.05 \; \forall 1 \leq i \leq k) = 1 - (1 - 0.05)^k \tag{2.2}$$

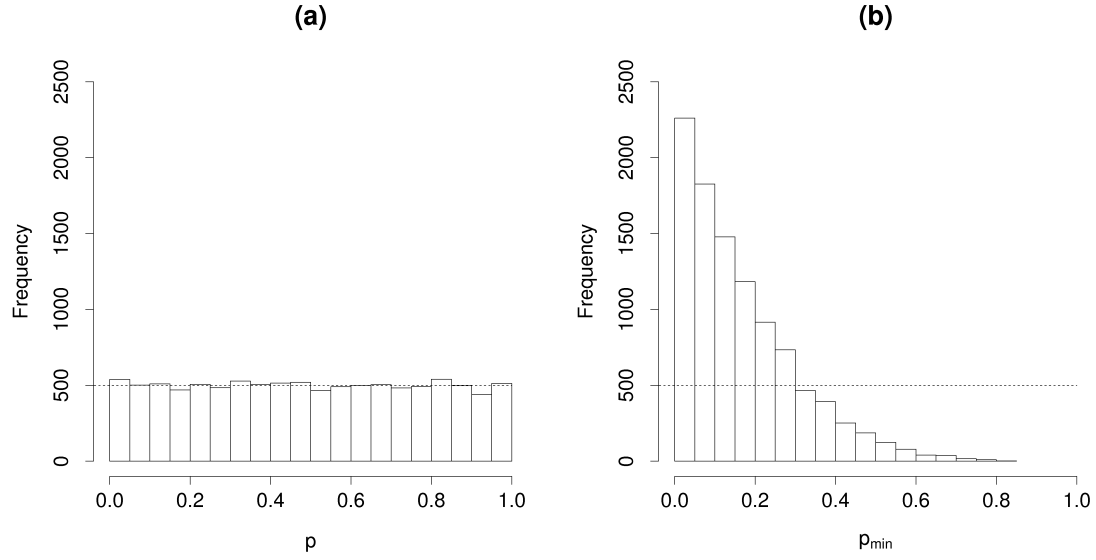

Figure 2.1: *Results of $N = 10,000$ simulations of two-sample t-tests based on normal data reflecting the null hypothesis with $n = 100$. The dashed line denotes the expected frequency of p-values assuming they are uniformly distributed under the null hypothesis. In **(a)**, this assumption is strikingly satisfied as the histogram bars are well-aligned with the expected level. Figure **(b)** shows frequencies of the minimum of 5 simulated p-values from $k = 5$ hypotheses tested each. Under the complete null hypothesis, the frequency of significant tests by far exceeds the nominal level $\alpha = 0.05$ as the actual type I error is $1 - (1 - 0.05)^5 = 0.23$. Note that the latter is still much higher for settings with many more then five tests; e.g. 0.98 for $k = 80$.*

This minimum is obviously not uniformly distributed under the complete null hypothesis as small values are more likely to occur. The probability for at least one of the k p-values to be in the interval $[0, 0.05)$ by chance is therefore much larger than 0.05; the probability density function (p.d.f.) of the minimum p-value distribution is derived from (2.2) as $f(p) = k(1 - p)^{k-1}$. For single Student t-tests as well as for the minimum of $k = 5$ such p-values, this density has been simulated and visualized in Figure 2.1. Tools are needed for modification of the tests to control the type I error by α also in multiple testing procedures.

Several possibilities to define error probabilities in multiple inference procedures have been discussed by Hochberg and Tamhane (1987). In (2.2), control of the probability for *at least one* type I error was intuitively required, which Tukey (1953) recommended

as the definition that "should be standard, rarely will any other be appropriate". In the literature, this is commonly referred to as the *familywise error* (FWE). Formally, the FWE can be further distinguished as follows:

(1) Partial familywise error (pFWE): the FWE is controlled in the strong sense, i.e. the condition

$$P\left(\exists i \in J : H_{0i} \text{ is rejected} \,|\, \mathbf{H}_0^J\right) \leq \alpha \tag{2.3}$$

holds for the intersection $\mathbf{H}_0^J$ of any possible collection of true null hypotheses $\{H_{0i}\}_{i\in J} \subset \mathcal{H}_0$. Note that this probability depends on the choice of $J \subset I$.

(2) Common familywise error (cFWE): the FWE is controlled in the weak sense, i.e. the condition

$$P\left(\exists i \in I : H_{0i} \text{ is rejected} \,|\, \mathbf{H}_0^I\right) \leq \alpha \tag{2.4}$$

holds for the *complete* null hypothesis $\mathbf{H}_0^I$ only.

The multiple p-values can be adjusted in a way that the distribution of their minimum is uniform on the interval $[0, 1]$ and the pFWE is controlled by the significance level α. If $K \subset I$ denotes the subset of true null hypotheses, this is attained by the definition $\tilde{p}_i := P\left(\max_{j\in K} |T_j| \geq |t_i| \,\middle|\, \mathbf{H}_0^K\right)$ of the adjusted *p*-value for hypothesis H_{0i}, extending the usual univariate p-value representation $p_i = P\left(|T_i| \geq |t_i| \,|\, H_{0i}\right)$. However, if for all subsets $J \subset K$, the joint distribution of the test statistics $\{T_j\}_{j\in J}$ does not depend on which of the hypotheses in $I \setminus J$ are true, the value of $\tilde{p}_i$ will be the same if the expression is considered for the whole set I instead:

$$\tilde{p}_i = P\left(\max_{j\in I} |T_j| \geq |t_i| \,\middle|\, \mathbf{H}_0^I\right) \tag{2.5}$$

This modification will be very helpful for practical use in the following chapters. Note that if different test statistics are used or both one- and two-tailed tests are among the family of tests, the definition (2.5) can lead to unbalanced multiplicity adjustment. This can be overcome using a definition based on the corresponding univariate p-values instead (Westfall and Young, 1993). As the current chapter is intended as a review of classical MCP, the historical notation will be used from now on, where inferences are given by critical points or confidence statements, respectively, rather than in terms of adjusted p-values. Without loss of generality, the notation for *two-sided* hypotheses as stated in (2.1) is used.

For the problem of pairwise comparisons in mean, the simultaneous confidence intervals

for the difference of population means μ_{i_1} and μ_{i_2} are of the form

$$\mu_{i_1} - \mu_{i_2} \in \left[\bar{\mathbf{X}}_{i_1} - \bar{\mathbf{X}}_{i_2} \pm \xi_i \sqrt{\frac{\hat{\sigma}_{i_1}^2}{n_{i_1}} + \frac{\hat{\sigma}_{i_2}^2}{n_{i_2}}} \right]$$

with critical points ξ_i and sample variance estimates $\hat{\sigma}_{i_1}^2$ and $\hat{\sigma}_{i_2}^2$. There are several reasons for a common choice $\xi_i \equiv \xi$ for these the most important of which is that calculation of the ξ_i is then much more convenient (Hochberg and Tamhane, 1987). The common critical point has to be chosen in a way that

$$P\left(\max_{i \in I} |T_i| \geq \xi \,\middle|\, \mathbf{H}_0^I \right) = \alpha \tag{2.6}$$

as rejecting at least one of the hypotheses in $\mathcal{H}_0$ is equivalent to $\max_{i \in I} |T_i| \notin R_0$, where the non-rejection region has got the form $R_0 = \left\{ \mathbf{t} \in \mathbb{R}^k | \|\mathbf{t}\|_\infty \leq \xi \right\}$ with $\| \cdot \|_\infty$ denoting the maximum norm. The set R_0 is a k-dimensional cube of edge length ξ, following from the properties of $\| \cdot \|_\infty$ known from standard calculus.
Now, the multiple problem of inference on $\mathcal{H}_0$ essentially reduces to the determination of ξ. In the literature, there have been many efforts on calculation or approximation of expressions of the form in (2.6) for particular settings. Some essential ideas are given in the following sections but these do by no means offer a complete overview.

2.1 Bonferroni and Šidák approach

An approach to multiple testing that is well-known to practitioners is the *Bonferroni method*, which means to test each hypothesis H_{0i} at level $\frac{\alpha}{k}$ instead of α, where k is the total number of hypotheses. The critical point for the ith test is then $t_{\nu_i}^{(1-\alpha/k)}$, the $(1-\alpha/k)$ point of the t-distribution with $\nu_i = n_{i_1} + n_{i_2} - 2$ degrees of freedom. As an upper bound to the cFWE, the *Bonferroni inequality*

$$P\left(\exists i \in I : H_{0i} \text{ is rejected} \,|\, \mathbf{H}_0^I\right) = P(\max_{i \in I} |T_i| \geq |t_{\nu_i}^{(1-\alpha/k)}| \,|\, \mathbf{H}_0^I) \leq \sum_{i \in I} P(|T_i| \geq |t_{\nu_i}^{(1-\alpha/k)}| \,|\, \mathbf{H}_0^I) \tag{2.7}$$

holds independently from the distribution of the T_i, but is a somewhat rough upper bound, contributing to the low power of the Bonferroni procedure. The familywise level is controlled because it follows from (2.7) that

$$P\left(\exists i \in I : H_{0i} \text{ is rejected} \,|\, \mathbf{H}_0^I\right) \leq k\frac{\alpha}{k} = \alpha \tag{2.8}$$

This requires the p-values involved in the analysis to be uniformly distributed on the interval $[0,1]$ which is true, for instance, whenever the Student t-test is applied to normally distributed data. However, the uniformity assumption is not always valid, e.g. in applications to non-normal populations or sparse binary data. In these cases, the Bonferroni method may yield extremely conservative decisions.
As another approach that is closely related to this, the *Šidák multiple testing method* offers a slight improvement compared to the Bonferroni method. The univariate hypotheses are tested at level $1-(1-\alpha)^{1/k}$ instead of α such that the critical point is $t_{\nu_i}^{(1-(1-\alpha)^{1/k})}$ for the i-th test. To show that the cFWE is still bounded by α, the Bonferroni inequality is not needed. Using the rules known from probability theory, it follows that

$$\begin{aligned} P\left(\exists i \in I : H_{0i} \text{ is rejected} \,|\, \mathbf{H}_0^I\right) &= P(\max_{i\in I} |T_i| \geq |t_{\nu_i}^{(1-(1-\alpha)^{1/k})}| \,|\, \mathbf{H}_0^I) \\ &= 1 - P(\forall i \in I : |T_i| < |t_{\nu_i}^{(1-(1-\alpha)^{1/k})}| \,|\, \mathbf{H}_0^I) \\ &= 1 - \prod_{i\in I} P(|T_i| < |t_{\nu_i}^{(1-(1-\alpha)^{1/k})}| \,|\, \mathbf{H}_0^I) \qquad (2.9) \\ &= 1 - ((1-\alpha)^{1/k})^k) = \alpha \end{aligned}$$

if the univariate p-values are assumed to be independent and uniformly distributed on $[0,1]$, i.e. the Šidák-adjusted p-values are *exact* in such a setting. Note that under certain conditions, (2.9) is still valid as a "$\leq$" relation if the independence assumption on the p-values is dropped (Šidák, 1967 and Jogdeo, 1977); i.e. the cFWE is controlled but the Šidák method is then no more exact. Nevertheless, the Šidák multiple testing method is in general more powerful than the Bonferroni procedure which can be proven by Taylor expansion of the levels α/k and $1-(1-\alpha)^{1/k}$.

2.2 Tukey's and Scheffé's method

Tukey (1953) proposed an approach to multiple inference where all pairwise comparisons are to be performed in a collection of m groups in which a normal distributed parameter is measured; i.e. the hypotheses of interest are of the form in (2.1) with $I = \{1,...,m\} \times \{1,...,m\}$. This involves $k = \binom{m}{2}$ single tests. For the balanced case $n_1 = ... = n_m \equiv n$, the statistic $\max_{i\in I} |T_i|$ is distributed as the *Studentized range random variable*

$$Q_{k,\nu} = \frac{\max_{1\leq i<j\leq k} |Z_i - Z_j|}{\sqrt{U/\nu}}, \qquad (2.10)$$

where U is χ^2_ν-distributed with $\nu = n-1$ degrees of freedom and $Z_1, \ldots, Z_k$ are independent standard normal variables. The critical point for the multiple procedure is then its

upper α-quantile $\xi := Q_{k,\nu}^{(\alpha)}$ that can be obtained from appropriate tables e.g. in the appendix of Hochberg and Tamhane (1987). The cumulative distribution function (c.d.f.) of the Studentized range random variable is also given there.
This is the classical *T-procedure* as it was initially introduced by Tukey (1953). For the balanced case, it has been shown by Gabriel (1970) that the corresponding simultaneous confidence intervals are the shortest among all intervals with constant length that keep the level $1-\alpha$. For general unbalanced designs, an approximate solution to (2.6) is classically given by the *Tukey-Kramer procedure* (Tukey, 1953 and Kramer, 1956): the upper α-point $Q_{k,\nu}^{(\alpha)}$ of the Studentized range distribution is taken as the critical point also in the unbalanced case. Hayter (1984) proved that the inequality

$$P\left(\mu_i - \mu_j \in \left[\bar{Y}_i - \bar{Y}_j \pm Q_{k,\nu}^{(\alpha)} S \sqrt{\tfrac{1}{2}\left(\tfrac{1}{n_i}+\tfrac{1}{n_j}\right)}\right]\right) \geq 1-\alpha$$

holds for arbitrary sample sizes $n_1, ..., n_m$. The test results from this approach are then no more exact but always keep the nominal significance level α. Hayter (1984) also showed that the procedure performs conservative for strongly unbalanced sample size allocation.

Another classical approach for multiple comparisons in general unbalanced designs is the multiple procedure of Scheffé (1953) that is commonly known as the *S-procedure*. The critical point for any collection of multiple comparisons is defined by $\xi := \sqrt{(m-1)F_{m-1,\nu}^{(\alpha)}}$, where $F_{m-1,\nu}^{(\alpha)}$ denotes the upper α-quantile of the multivariate F-distribution with parameters $m-1$ and ν. This procedure is exact for the case of all real-valued contrasts being tested, i.e. the identity

$$P\left(\forall \mathbf{c} \in \mathbb{C}^m : \sum_{i=1}^m c_i \mu_i \in \left[\sum_{i=1}^m c_i \bar{\mathbf{X}}_i \pm \hat{\sigma}\sqrt{(m-1)F_{m-1,\nu}^{(\alpha)}}\left(\sum_{i=1}^m \frac{c_i^2}{n_i}\right)^{1/2}\right]\right) = 1-\alpha \quad (2.11)$$

holds for any choice of m with the linear space $\mathbb{C}^m$ defined previously, i.e. the number of tests is infinite and even *uncountable* in contrast to the prior discussion. The proof of (2.11) is based on the fact that

$$\sum_{i=1}^m \frac{\bar{\mathbf{X}}_i^2}{\sigma^2/n_i} \sim \chi_{m-1}^2 \quad \text{and} \quad \frac{\nu\hat{\sigma}^2}{\sigma^2} \sim \chi_\nu^2$$

under the complete null hypothesis; hence it follows that $\frac{\nu}{m-1}$ times the ratio of both statistics is $F_{m-1,\,\nu}$-distributed. Details on this are given in Hsu (1996) and the classical paper of Scheffé (1953).
The S-procedure provides a tool for a wide range of multiple contrast tests, but for particular subsets of hypotheses, the performance can be very poor. As a corollary to (2.11),

it follows that the coverage probability is greater than $1-\alpha$ for particular contrast matrices: for the all pairwise comparisons problem, the intervals are in fact wider than if the Tukey-Kramer procedure is applied (Hsu, 1996).

2.3 Multiple inferences using the multivariate *t*-distribution

The Bonferroni and Šidák approaches to the multiple problem $\mathcal{H}_0$ assume that the k test statistics involved in the analysis are stochastically independent. This is accompanied by an unreasonable waste of information: consider the marginal case where the same null hypothesis is tested k times, using always the same test, i.e. the correlation matrix of the test statistics is a $k \times k$ matrix containing the value 1 in each cell. Obviously no multiplicity adjustment is needed in such a setting as it is equivalent to the univariate test. Now, speaking heuristically, if the statistics from two different tests in the analysis are highly correlated, indicating that the hypothesis of the first test is associated with that of the second, there is less multiplicity adjustment necessary than for two independent test statistics. In general, if any two of the k test statistics in the analysis have non-vanishing correlations or if the correlation matrix of the multiple problem is any other than the identity matrix, this information may be used to construct more powerful tests. Modifications allowing for more general MCP exist for the T-procedure. These methods are no longer needed as solutions of (2.6) can nowadays be calculated numerically on a standard desktop computer: for a multiple procedure, the multivariate distribution of the vector $\mathbf{T}$ can be involved in the determination of a rejection region $R = \mathbb{R}^k \setminus R_0 = \left\{\mathbf{t} \in \mathbb{R}^k | \|\mathbf{t}\|_\infty > \xi\right\}$, where the critical point ξ is to be chosen in a way that (2.6) holds. For multiple t-tests, the value of ξ is uniquely determined by the condition

$$\frac{\Gamma\left(\frac{k+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{|\mathbf{R}|(\nu\pi)^k}} \int_R \left(1 + \frac{\mathbf{t}^T\mathbf{R}^{-1}\mathbf{t}}{\nu}\right)^{-\frac{k+\nu}{2}} d\mathbf{t} \;=\; \alpha,$$

where the integrand represents the p.d.f. of the centered multivariate t-distribution with correlation matrix $\mathbf{R}$ and ν degrees of freedom. From this, a value of ξ that satisfies condition (2.6) can in principle be calculated numerically. This may be difficult and connected to extremely high computational effort. New algorithms for this purpose have been developed and discussed by Genz and Bretz (1999 and 2002) and Bretz, Genz and Hothorn (2001), implementations of which are now available in statistics software packages like R and SAS. These methods will be applied when comparing the coverage of the resampling-based intervals to classical methods.

2.4 Step-down multiple comparison procedures

Stepwise MCP generally comprise *step-down* and *step-up* methods. The major part of literature focuses on the step-down approach as the same theory applies to step-up methods analogously. These are therefore omitted in this section.
The union of rejection regions for the univariate tests with $i \in J$ is the rejection region of the test on the intersection null hypothesis $\mathbf{H}_0^J$ (union-intersection method, Roy 1953). A step-down procedure for the family $\mathcal{H}_0$ of null hypotheses typically begins with a level α test on the complete null hypothesis $\mathbf{H}_0^I$ to make inference on the question if *any* hypothesis in the family $\mathcal{H}_0$ can be rejected. If this test fails to reject $\mathbf{H}_0^I$, the procedure stops at this point and all hypotheses H_{0i}, $i \in I$, are retained. Otherwise, all intersections $\mathbf{H}_0^J$ with $|J| = k - 1$ are tested at level α and the implied hypotheses are all retained if the test on $\mathbf{H}_0^J$ is not significant. Subsequently, the algorithm proceeds stepping through the hierarchy of hypotheses in the same fashion down to the elementary hypotheses. More formally, a step-down procedure can be constructed by the *closed test principle* proposed by Marcus, Peritz and Gabriel (1976). The *closure* of $\mathcal{H}_0$ is formed by the system of all possible intersections, i.e.

$$\bar{\mathcal{H}}_0 := \{\cap_{i \in J} H_{i0} \,|\, J \subset I\} \tag{2.12}$$

Now, a hypothesis $\mathbf{H}_0^K$ is retained if any null hypothesis $\mathbf{H}_0^J$ with $J \supset K$ is retained; otherwise $\mathbf{H}_0^K$ is tested at the unadjusted level α. This approach has been shown to control the pFWE by Marcus, Peritz and Gabriel (1976).
The procedure can be displayed in a clearly arranged system as shown in Figure 2.2 for the $k = 3$ case. This is much handier as compared to larger values of k where a tree structure like in Figure 2.2 can become very complicated: the number of tests to perform is $\binom{k}{2}$ for the all pairwise comparisons problem and therefore increasing quadratically. Some considerations on optimization are therefore reasonable to find some kind of shortcut for the procedure.
Holm (1979) proposed a *sequentially rejective* MCP where first the observed values of the test statistics are arranged in increasing order $t_{(1)}, ..., t_{(k)}$ with the corresponding hypotheses $H_{0(1)}, ..., H_{0(k)}$, where rejection of the complete null hypothesis is equivalent with $H_{0(k)}$ being rejected. The multiple hypotheses problem is then reduced to a collection of $k - 1$ hypotheses. Proceeding in the same way for all subset hypotheses, the critical points are obtained as a monotonically decreasing sequence

$$\xi_k \geq \xi_{k-1} \geq ... \geq \xi_1, \tag{2.13}$$

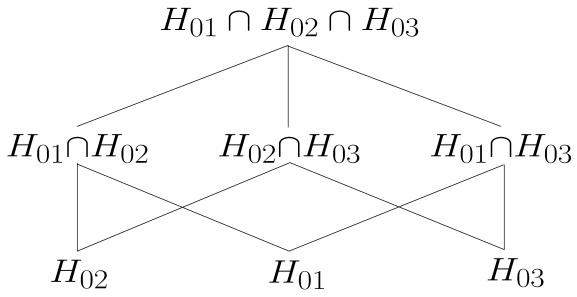

Figure 2.2: *Closed test principle. If $H_{01} \cap H_{02} \cap H_{03}$ is rejected, the hypotheses $H_{01} \cap H_{02}$, $H_{02} \cap H_{03}$ and $H_{01} \cap H_{03}$ are tested. The elementary hypotheses, i.e. H_{01}, H_{02} or H_{03}, are tested only if the respective two hypotheses which imply these can be rejected. The procedure can be proven to control the nominal level α.*

depending on the correlation structure among the test statistics. Now, if the hypotheses $H_{0(k)}, ..., H_{0(j+1)}$ are all rejected but $H_{0(j)}$ is retained, the procedure is stopped and the remaining hypotheses $H_{0(j-1)}, ..., H_{0(1)}$ are also retained. From (2.13), it immediately follows that a step-down procedure generally provides a greater power than the corresponding single-step method because the critical constants resulting from the stepwise algorithm are always equal or smaller than those based on a single-step method.
Obviously, the choice of the critical point sequence is crucial for the performance of the sequentially rejective procedure. It can be based on the Bonferroni or Šidák method which are known to be conservative, essentially because any distributional knowledge on the data is ignored. As shown for the single-step approach, the intersection hypotheses $\mathbf{H}_0^J$ can be tested by appropriate F-tests or Studentized range tests (Hochberg and Tamhane, 1987). The α-points of the F-test are obtained as $F_{|J|,\nu}^{(1-\alpha)}$ for a particular subset J of hypotheses. In principle, the closed test can be based on all particular testing techniques; also the resampling-based approach given in the following section is convertable to a step-down procedure. This has been discussed by Westfall and Young (1993).

2.5 Resampling-based approach

The method based on the multivariate c.d.f. of the test statistics fails for non-normal and especially for heteroscedastic data. The correlation structure and distributional shape should be involved in the analysis also if the joined c.d.f. of the test statistics is not the

multivariate t-distribution but any other c.d.f. which has unknown correlation structure or is completely unknown. A resampling-based approach has got the desirable features that on the one hand, it takes into account all the distributional information contained in the data and on the other hand is not too complex on an analytical level.
Westfall and Young (1993) presented a comprehensive overview of resampling methods in multiple hypotheses problems. Basically, the idea is to represent the distribution of the test statistics under the null hypothesis by repeatedly simulating data sets from the observations in a way that the null is reflected. From these data, the statistics are re-calculated and their empirical c.d.f. is used as the reference distribution for the test decision. In multiple problems, reflection of the null hypothesis needs further discussion: prior to testing, the particular subcollection of true hypotheses is, of course, unknown as otherwise there would be no reason to perform any test. The condition under which strong control of the familywise error is provided if resampling is done under the *complete* null hypothesis $\mathbf{H}_0^I$ is restated in the next section.
The distribution of the test statistic is simulated by recalculating it from data sets that are repeatedly generated from the original data. Typically, the latter are modified such that they satisfy the null hypothesis, e.g. by centering to a common mean. If hypotheses on the population means, e.g. $\mu_1 = \mu_2$ and $\mu_1 = \mu_3$ are tested by t-tests on the respective (unpaired) data vectors $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$, the correlations of the test statistics are determined by the proportion of sample sizes in the commonly and not commonly used groups; that is to say, in a balanced design, the correlation between the two statistics is exactly 0.5. In a resampling-based approach, this correlation is involved automatically in a way that the observations $\mathbf{X}_1$ are used for both tests.
Consider a continuous data application where the normality assumption must be dropped. In the one-sample case, even a single test statistic is then not exactly t-distributed as normality is required in the definition of the t-distribution. For more than one hypothesis, the c.d.f. of the maximum test statistic is involved which has a somewhat skewed shape even for the normally distributed case. This skewness becomes very serious if in addition the underlying data origin from a non-symmetric population. Westfall and Young (1993) report this in detail for lognormal data in a multiple setting. However, these problems are much less important for two-sample comparisons as the corresponding statistics are approximately normal in particular for balanced designs and a similar skewness shape in both groups. Nevertheless, the maximum distribution is still non-normal and is approximated better by a resampling-based approach.
If on the other hand, the equality of the variances over the groups cannot be supposed, the test can be based on the Satterthwaite test statistic instead of the classical t-statistic

which was implicitly introduced in Chapter 3. This statistic is no more t-distributed and the c.d.f. has to be approximated in an appropriate sense, which is commonly known as the *Behrens-Fisher problem*. Many proposals for its solution exist in the literature. The problem does not occur if the c.d.f. are obtained by resampling as the deviation from the t-distribution is then also present in all statistics based on the resampled data.
Resampling-based methods are of considerable interest for special test statistics where the derivation of the c.d.f. is not feasible or analytically intractable. Examples of this are the the min-test (Laska and Meisner, 1989) and the AVE- and MAX-statistics proposed by Hung, Chi and Lipicky (1993) and Hung (2000) that are frequently applied to the evaluation of drugs with multiple compounds (Chapter 3). The theory of bifactorial designs developed by Hung and others lacks for a general approach with arbitrary distributions underlying the data, e.g. for cases where non-normal or heteroscedastic data are involved. A resampling-based approach to the problem is suitable to avoid the calculations of c.d.f.'s for the multiple hypotheses problem associated with factorial designs as well as for the AVE- and MAX-test. The theory and performance of this application are developed in the following chapters, pointing out some special problems that arise.
A general decision prior to the application of resampling is whether permutation- or bootstrap-methods should be used. Different models and perceptions of the null hypotheses underlie these approaches.

2.5.1 Bootstrap methods

The null hypothesis can be reflected by centering the data involved in the respective test to a common mean and resampling with replacement from these data. This is what the term *bootstrap* is commonly used for (Efron, 1979), paralleling the iterative re-use of data to lifting yourself by your own bootstrap. General guidelines for bootstrap hypothesis testing were given by Hall and Wilson (1991):

(A1) "Care should be taken to ensure that even if the data might be drawn from a population that fails to satisfy H_0, resampling is done in a way that reflects H_0."

(A2) "Bootstrap hypothesis testing should use methods that are already recognized as having good features in the closely related problem of confidence interval construction". That is, the statistic should be pivotal which means that under the null, their distributions should not depend on which distribution generated the data.

Babu and Singh (1983) showed that bootstrap resampling on pivotal statistics has got better convergence properties. Summarizing both rules for the t-test, it is first impor-

tant to center the data with the *sample mean* instead of the supposed population mean because the statistic then reflects the null hypothesis even if it is not true. Second, the resampled statistic has to be studentized dividing by the scale parameter $\hat{\sigma}/\sqrt{n}$ to make it a pivotal statistic. Westfall and Young (1993) gave an extension of these guidelines for *multiple* bootstrap hypothesis testing:

(B1) The guidelines of Hall and Wilson hold for all marginal distributions in the multiple setting.

(B2) Resampling has to be done in a way that reflects the *complete* null hypothesis.

(B3) The *subset pivotality condition* holds: for all subsets $J \subset I$ of true null hypotheses, the joint distribution of $\{T_i\}_{i \in J}$ is identical under the restrictions $\mathbf{H}_0^J$ and $\mathbf{H}_0^I$.

The latter condition allows to resample the data under the complete null hypothesis $\mathbf{H}_0^I$ instead of partial null hypotheses which would be impossible as the subset of *true* null hypotheses is unknown. Westfall and Young (1993) showed that the partial familywise error (pFWE) is protected by α if resampling is done under the complete null hypothesis and the subset pivotality condition holds for the tests. An outline for an algorithm to calculate adjusted p-values for the family $\mathcal{H}_0$ is now given.

Algorithm 2.1 *1. Calculate the statistics T_i from the data $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$. Initialize counting variables $z_i = 0$ for $i \in I$.*

2. Generate data sets $\mathbf{X}_{i_1}^*$ and $\mathbf{X}_{i_2}^*$ by drawing with replacement samples from the centered versions of $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$. Calculate the statistics T_i^* of interest from these for all $i \in I$.

3. If $\max_{i \in I} |T_i^*| \geq |T_i|$, increase the corresponding counting variable by one: $z_i = z_i + 1$.

4. Repeat (2) and (3) N times. The estimated value of $\tilde{p}_i$ is then $\tilde{p}_i^{(N)} = z_i/N$.

The guidelines (A1) and (A2) are kept by Algorithm 2.1 as well as (B1), (B2) and (B3). Reflection of $\mathbf{H}_0^I$ is achieved by centering all data to the common mean 0. The underlying statistics are pivotal as they are studentized by an estimate $\hat{\sigma}/\sqrt{n}$ of the standard error. The subset pivotality condition holds because for a subset $J \subset I$ of true null hypotheses, the $\{T_i\}_{i \in J}$ are multivariate *t*-distributed with a certain correlation matrix that does not depend on which particular superset $K \supset J$ of null hypotheses is true.

Another application of the bootstrap is the calculation of confidence intervals for the differences $\mu_{i_1} - \mu_{i_2}$ for all $(i_1, i_2) \in I$. The intervals $\{C_i\}_{i \in I}$ have *simultaneous* coverage probability $1 - \alpha$ if

$$P\left(\bigcap_{i \in I} \{\mu_{i_1} - \mu_{i_2} \in C_i\}\right) = 1 - \alpha,$$

i.e. the probability for *all* intervals (instead of each single interval) to contain the respective parameters is $1 - \alpha$.
Construction of confidence intervals requires an estimate for the critical point ξ of the test statistic. In the multiple case, a common value of ξ can be chosen according to (2.6) and some possibilities for calculation of ξ have been discussed. As a convenient approach for general data, the critical point can also be obtained by a bootstrap algorithm as proposed by Edwards and Berry (1987). All samples are centered by zero prior to the procedure, i.e. resampling is done under the complete null hypothesis $\mathbf{H}_0^I$. The empirical $(1 - \alpha)$-quantile of $\max_{i \in I} |T_i|$ is taken as an estimate of ξ. For the family $\mathcal{H}_0$ of hypotheses, simultaneous confidence intervals are obtained by the following algorithm.

Algorithm 2.2 *1. Generate data sets $\mathbf{X}_{i_1}^*$ and $\mathbf{X}_{i_2}^*$ by drawing with replacement samples from the centered versions of $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$.*

2. Compute the statistics T_i^* of interest from these for all $i \in I$ and store the value $\max_{i \in I} |T_i^*|$.

3. Repeat steps (2) and (3) N times. The m-th order statistic $W_{(m)}$ of the N values of $\max_{i \in I} |T_\gamma^{j*}|$ with $m = [(N + 1)(1 - \alpha)]$ is an estimate of the critical value ξ.

The consistency of the bootstrap estimate is based on the fact that the m-th order statistic converges to the $(1 - \alpha)$-quantile of the underlying population as $n \to \infty$. A proof of this statement is given, for instance, by Shao (1999, p. 305).
There are a few limitations of Algorithm 2.2. First, this approach will always result in symmetric intervals as the critical point is the same for both confidence bounds. Furthermore, using the two-sample t-statistic is valid only in the case of normal distributed data in both groups which is, however, less important for large samples and a similar skewness in both groups. To overcome these problems, an alternative approach to bootstrap-based confidence intervals proposed by Efron and Tibshirani (1993) may be suitable. In the univariate case, the $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ percentiles of the bootstrap sample mean distribution are used as the lower and upper limits for the confidence interval of the population

mean, whereas there is currently no extension of this method to multiple hypotheses problems. However, as the percentile methods makes immediate use of the empirical distribution of the parameter estimate, it applies to a large variety of distributional situations. The following algorithm forms an outline for simultaneous confidence intervals using the percentile method. Basically, the idea is that a bootstrap iteration underlying an extreme value for a particular parameter estimate will lead to extreme values also for estimates of other parameters if there is a non-vanishing correlation.

Algorithm 2.3 *1. For $i \in I$, generate data sets $\mathbf{X}_{i_1}^*$ and $\mathbf{X}_{i_2}^*$ by drawing with replacement samples from $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$.*

2. Calculate the mean difference $\bar{\mathbf{X}}_{i_1} - \bar{\mathbf{X}}_{i_2}$ for all $i \in I$ and append these values as a row to a matrix M. Repeat steps (1) and (2) N times.

3. For $j = 1, ..., k$, eliminate the rows of M with the smallest and largest value in the jth column. Repeat this step $\frac{\alpha N}{2k}$ times.

4. For the jth comparison, the estimates for the confidence limits are $\min_{1 \le i \le k} M_{ij}$ and $\max_{1 \le i \le k} M_{ij}$.

The most extreme $\frac{\alpha N}{2k}$ iterations are excluded not only from the empirical distribution of each sample mean difference, which would be equivalent to the Bonferroni method, but the iterations yielding the most extreme values in column j are also omitted from the distributions of the remaining $k - 1$ comparisons, making implicit use of the correlation structure. Thus, the width of the intervals based on Algorithm 2.3 is uniformly smaller or equals that of Bonferroni intervals.

2.5.2 Permutation-based resampling

In contrast to the preceding sections, the family of hypotheses $\mathcal{H}_0$ is now assumed to have a structure representing equality of the respective c.d.f., that is to say $F_{i_1} = F_{i_2}$ vs. $F_{i_1} \neq F_{i_2}$ for $i = (i_1, i_2) \in I$ instead of the parametric statements given in (2.1). This is equivalent to the *rerandomization null hypothesis* that the data in the respective two groups are a random rearrangement of the pooled sample $(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$. A p-value that is *exact* given the observed samples can therefore be generated by calculating the test statistics from all $\binom{n_{i_1}+n_{i_2}}{n_{i_1}}$ possible allocations of these data in two groups. As the order

within the groups is not relevant for calculation of the test statistics, the exact p-value is the fraction

$$\tilde{p}_i = \frac{\#\{\max_{i\in I} |T_i^*| \geq |T_i|\}}{\binom{n_{i_1}+n_{i_2}}{n_{i_1}}}, \tag{2.14}$$

where the numerator can be obtained by evaluation of the test statistics for all possible allocations of the data to the groups. However, for applications with even moderate large data amount, this requires heavy computational effort which exceeds technical limits: the number of possible allocations of the data to a balanced two-sample design would be $\binom{n_1+n_2}{n_1} = \binom{2n}{n}$, which is obviously not computationally tractable for, say, $n > 10$. Instead, the *exact* probability given in (2.14) can be closely approximated by evaluation of a large number N of possible permutations. This is attained by repeatedly shuffling the observed data from the pooled treatment groups. As the null hypothesis is then reflected by the new samples, the reference distribution for the test can be determined by repeatedly recalculating the test statistics from those. The following algorithm is approriate for testing a family $\mathcal{H}_0$ of hypotheses, where the structure is like in (2.1), but equality of the c.d.f. is explored as mentioned above.

Algorithm 2.4

1. *Calculate the statistics T_i from the data $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$. Initialize counting variables $z_i = 0$ for $i \in I$.*
2. *Generate data sets $\mathbf{X}_{i_1}^*$ and $\mathbf{X}_{i_2}^*$ by drawing without replacement samples from the pooled data $(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$. Calculate the statistics T_i^* of interest from these for all $i \in I$.*
3. *If $\max_{i\in I} |T_i^*| \geq |T_i|$, increase the corresponding counting variable by one: $z_i = z_i + 1$.*
4. *Repeat 2-3 N times. The estimated value of $\tilde{p}_i$ is then $\tilde{p}_i^{(N)} = z_i/N$.*

Note the close similarity between Algorithm 2.1 and 2.4.

2.5.3 Relationship of permutation and bootstrap resampling

Technically, the major difference between bootstrap and permutation resampling is that in the former method, resampling is done *with replacement*, where resampling *without replacement* from the pooled samples is used by the latter. The second guideline for the bootstrap given by Hall and Wilson (1991) suggests that the statistic which the bootstrap procedure is based on should be pivotal. Efron and Tibshirani (1993) point out that in settings where both methods apply, in fact the bootstrap performs slightly more powerful

when using studentized statistics, whereas the permutation test is not affected by this. The results of both approaches are quite similar anyway and thus the distinction is only slightly relevant for practice. Nevertheless, some thoughts on the relationship between both are helpful to get a deeper understanding.

The rerandomization null hypothesis implies that all permutations of the data in the respective groups are equally likely to occur. This symmetry limits the permutation method to tests on hypotheses of the form $F_1 = F_2$, where F_1 and F_2 are two c.d.f. to be tested for equality. In particular, no parametric hypotheses, e.g. for equality of means or variances, can be tested. The range of applications is more general for the bootstrap as the above symmetry is not implied by parametric hypotheses. On the other hand, the assumption of equal probabilities for all permutations is also the fact where the *exactness* of the permutation methods given the observed data origins from (Section 2.5.2). In contrast to this, the bootstrap p-values are by no means exact; quite the contrary must there be a certain minimum of sample size to produce reasonably accurate results. Efron and Tibshirani (1993) state that "bootstrap methods are more widely applicable but less accurate," briefly summarizing the essential hallmarks of both permutation and bootstrap resampling.

Consider a binary data application where for $i = (i_1, i_2) \in I$, the binomial event rates π_{i_1} and π_{i_2} are compared between two binomial samples $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$. Strictly speaking, the number of events is a hypergeometric random variable but can be closely approximated by the binomial c.d.f. for large populations due to the asymptotic properties of the hypergeometric distribution. If the null hypothesis $H_{0i} : \pi_{i_1} = \pi_{i_2}$ is reflected by two bootstrap data vectors $\mathbf{X}^*_{i_1}$ and $\mathbf{X}^*_{i_2}$ of the same length $n = n_{i_1} = n_{i_2}$ which are drawn with replacement from the pooled sample $(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$, both of these have got the estimated event rate $\hat{\pi}_i = \frac{1}{2}(\hat{\pi}_{i_1} + \hat{\pi}_{i_2})$. From the new samples, a test statistic T_i is then calculated. Using the denotations $Z^*_{i_1} = \sum \mathbf{X}_{i_1}$ and $Z^*_{i_2} = \sum \mathbf{X}_{i_2}$, the probability that T_i has got the observed or an even more extreme value under H_{0i} is obtained by summing the expressions

$$P\left((Z^*_{i_1}, Z^*_{i_2}) = (k_{i_1}, k_{i_2})|\hat{\pi}\right) = \binom{n_{i_1}}{k_{i_1}} \hat{\pi}^{k_{i_1}} (1-\hat{\pi})^{n_{i_1}-k_{i_1}} + \binom{n_{i_2}}{k_{i_2}} \hat{\pi}^{k_{i_2}} (1-\hat{\pi})^{n_{i_2}-k_{i_2}}$$

over all values of (k_{i_1}, k_{i_2}) yielding a statistic at least as large as the observed value. Obviously, the p-value depends on the particular choice of T_i. If, on the other hand, the permutation test is applied, the value $k_{i_1} + k_{i_2}$ is equal for all possible permutations and the probability of the test statistic to be more extreme than the observed value equals the probability of a more extreme distribution of the events to the integers k_{i_1} and k_{i_2}. Thus, the procedure always results in Fisher's exact test, regardless which statistic it is based

on, and is therefore conservative as compared to the bootstrap approach: for binary data, simulation results of Westfall and Young (1993) showed that the bootstrap exploits the nominal level α better than the permutation methods; they propose a *with replacement* resampling algorithm as a generally more powerful approach for binary data.

2.5.4 Performance of the resampling-based approach

For the bootstrap as well as for permutation-based resampling, errors essentially arise for three reasons. First, random number generation for the decision on which bootstrap samples to choose is never perfect, but the deviation resulting from this is negligible because pseudo random numbers can be computed with an extremely high precision on modern machines. Second, consider the simulation standard error that depends on the number of bootstrap iterations. The variables z_i counting the frequency of the event $\max_{i\in I}|T_i^*| \geq |T_i|$ in Algorithms 2.1 and 2.4 are binomial distributed with parameters N and $\tilde{p}_i$ for each combination $i \in I$ and therefore have the probability density function

$$f^{z_i}(k) = \binom{N}{k}(1-\tilde{p}_i)^{N-k}\tilde{p}_i^k,$$

i.e. the simulation standard error of the estimate $\hat{\tilde{p}}$ for the adjusted p-value has got the representation

$$\text{se}[\tilde{p}_i^{(N)}] = \sqrt{\frac{\text{Var}[z_i]}{N^2}} = \frac{(1-\tilde{p}_i)\tilde{p}_i}{N} = O(N)$$

and hence can be brought arbitrarily close to zero if sufficient computational resources are available.

For the confidence intervals, let G be the c.d.f. of the statistic $\max_{i\in I}|T_i|$. With the denotation $m = [(N+1)(1-\alpha)]$, the deviation from the specified level α induced by simulation is given by $G(W_{(m)})$, where $W_{(m)}$ is the m-th order statistic of the N values of $\max_{i\in I}|T_i^*|$ based on the respective N bootstrap samples and $G(W_{(m)}) = 1-\alpha$ represents a zero deviation. Because $G(W_{(m)})$ is beta-distributed with parameters m and $N-m+1$ (Edwards and Berry, 1987), the simulation standard error for the intervals is $\text{se}[G(W_{(m)})] = \frac{\alpha(1-\alpha)}{N}$ and therefore of the same order as above. For reasonable allocation of resources, the algorithms should stop as soon as N is large enough for the standard errors to satisfy the condition $\text{se}[\tilde{p}_i^{(N)}] \leq \varepsilon$ or $\text{se}[G(W_{(m)})] \leq \varepsilon$, respectively, where ε is a prespecified error bound.

A third important kind of error results from the use of $\hat{G}_n$ instead of G and therefore cannot be decreased with computational power. Anyway, following from the Glivenko-Cantelli theorem, $\sup_{x\in\mathbb{R}}\left|\hat{G}_n(x) - G(x)\right| \to 0$ holds for $n \to \infty$ if G is continuous and

the observations are independent. In particular, this shows that any kind of resampling-based calculations is only reasonable for data samples that are not extremely small; on the other hand, the third error source is negligible for large-sample applications since the empirical c.d.f. approximates the population well in these cases. Singh (1981) and Babu and Singh (1983) discussed convergence rates of bootstrap estimates for increasing sample sizes.

In the simulation experiments used for evaluation of algorithms in the following chapters, the probability $\alpha_0 = P(p \leq \alpha)$ will be estimated. As pointed out by Westfall and Young (1993), the estimate $\hat{\alpha}_0$ matches the true value closer if more computational power is allocated to the outer than to the inner loop. As the total simulation size is limited by the availability of system resources, $25,000$ simulations with $N = 15,000$ bootstrap iterations are a reasonable choice. This yields a 0.95 confidence interval of $\hat{\alpha}_0 \pm 1.96\sqrt{\frac{\hat{\alpha}_0(1-\hat{\alpha}_0)}{25,000}}$ for the true value of α_0, e.g. 0.05 ± 0.0027 for $\hat{\alpha}_0 = 0.05$. When interpreting the simulation results, it should therefore be kept in mind that there is still an uncertainty in the third decimal place. Using these settings and a `C++` implementation for the simulations in Chapters 4 and 5, e.g. those in Table 4.1, every single simulation for the largest sample size $n = 250$ required several days of computation time on a 2 x 2.8 GHz Intel Pentium D CPU.

3 Efficacy analysis using the min-test

In the context of clinical trials on dose combinations, treatment with each component drug is understood as a factor with stages according to the doses applied, establishing a certain dose-response relationship. Bifactorial trial designs are used to test for the efficacy of combinations of two component therapies A and B. Each pair (i, j) identifies a unique dose combination on the grid $\mathbb{G}_0 = \{0, ..., A\} \times \{0, ..., B\}$, i.e. drug A can be applied in doses 0 (placebo) to A, analogous for drug B. The question of interest is if the respective groups $(i, j) \in \mathbb{G}$ have a significantly higher response than both mono therapy groups $(i, 0)$ and $(0, j)$ for the set $\mathbb{G} = \mathbb{G}_0 \setminus \{(i, j)|i = 0 \vee j = 0\}$.
For a pair (i, j), let $\mathbf{X}_{ij}$ denote a response vector with n_{ij} values from the respective treatment group. If μ_{ij} is the actual mean response value, it is reasonable to use the model assumption $\mathbf{X}_{ij} = \mu_{ij} + \varepsilon_{ij}$, where the errors ε_{ij} must not necessarily be identically distributed. In particular, they can be non-normal or the variances might be heterogeneous over the treatment groups. The associated (one-sided) test problem is

$$H_0^{ij} : (\mu_{ij} \leq \mu_{i0}) \vee (\mu_{ij} \leq \mu_{0j}) \quad \text{vs.} \quad H^{ij} : (\mu_{ij} > \mu_{i0}) \wedge (\mu_{ij} > \mu_{0j}). \tag{3.1}$$

The decision is based on the so-called min-test statistic $T_{ij}^{min} = \min\left\{T_{ij}^A, T_{ij}^B\right\}$ (Laska and Meisner, 1989), where

$$T_{ij}^A = \frac{\bar{\mathbf{X}}_{ij} - \bar{\mathbf{X}}_{i0}}{\sqrt{\frac{\hat{\sigma}_{ij}^2}{n_{ij}} + \frac{\hat{\sigma}_{i0}^2}{n_{i0}}}} \quad \text{and} \quad T_{ij}^B = \frac{\bar{\mathbf{X}}_{ij} - \bar{\mathbf{X}}_{0j}}{\sqrt{\frac{\hat{\sigma}_{ij}^2}{n_{ij}} + \frac{\hat{\sigma}_{0j}^2}{n_{0j}}}} \tag{3.2}$$

with the estimate for the population variance σ_{ij}^2 in the respective treatment groups denoted by $\hat{\sigma}_{ij}^2$. The idea is that the observed value of T_{ij}^{min} is equal or greater than a given critical value if and only if this is the case for *both* T_{ij}^A and T_{ij}^B. In the balanced and homoscedastic case, the min-statistic can equivalently be determined from the representation

$$T_{ij}^{min} = \frac{\bar{\mathbf{X}}_{ij} - \max\{\bar{\mathbf{X}}_{i0}, \bar{\mathbf{X}}_{0j}\}}{\hat{\sigma}/\sqrt{n}}.$$

According to the treatment groups in $\mathbb{G}$, various test statistics of the min-test type are calculated and the decision to reject or not to reject the null hypothesis (3.1) is made for each pair (i, j).

Note that (3.1) can equivalently be represented by $H_0^{ij} : \vartheta_{ij} \leq 0$ with $\vartheta_{ij} := \mu_{ij} - \max\{\mu_{i0}, \mu_{0j}\}$. The distribution of the min-statistic then depends on ϑ_{ij}, the parameter of interest, as well as on the *unknown* nuisance parameter $\delta_{ij} := \mu_{i0} - \mu_{0j}$ because the latter determines which of the two mean differences $\mu_{ij} - \mu_{i0}$ and $\mu_{ij} - \mu_{0j}$ is smaller and hence will influence the probabilities for $T_{ij}^{min} = T_{ij}^A$ and $T_{ij}^{min} = T_{ij}^B$. The c.d.f. of T_{ij}^{min} and its correlation structure is therefore a function of ϑ_{ij} and δ_{ij}.
As a way out of this problem, Hung (1993) proposed an α-level two-stage design, where the hypothesis $H_0 : \mu_{i0} = \mu_{0j}$ is tested in a first step. If it can be rejected, the min test is conducted in the second step, otherwise the pooled test $\frac{T_{ij}^A + T_{ij}^B}{2}$. Adjustment of the resulting p-values is necessary to protect the type I error level α. This contributes to the fact that this two-stage test is not very powerful as compared to other approaches (Hung, 1993).

3.1 Testing for the existence of efficacious dose combinations

The question if *any* combination is better than both of its components can be expressed by the global test problem

$$\begin{aligned} H_0 &: \forall (i,j) \in \mathbb{G} : (\mu_{ij} \leq \mu_{i0}) \vee (\mu_{ij} \leq \mu_{0j}) \\ \text{vs.} \quad H &: \exists (i,j) \in \mathbb{G} : (\mu_{ij} > \mu_{i0}) \wedge (\mu_{ij} > \mu_{0j}). \end{aligned} \tag{3.3}$$

As proposed by Hung, Chi and Lipicky (1993), inference can be based on the AVE-test $T_{ave} = (AB)^{-1} \sum_{i=1}^{A} \sum_{j=1}^{B} T_{ij}^{min}$ or the MAX-test $T_{max} = \max_{(i,j)\in\mathbb{G}} T_{ij}^{min}$. The respective distributions were given by Hung, Chi and Lipicky (1993) under the restriction that the data are normally distributed with a common value σ^2 for the variance and balanced sample size allocation with n individuals in each group. They depend on the primary parameters ϑ_{ij} as well as on the nuisance parameters δ_{ij} that are *unknown* in all practically relevant settings.
The power functions of T_{ave} and T_{max} are closely related to their respective c.d.f.. For monotonicity reasons, the type I error of the AVE-test is bounded by the significance level α which is obtained by taking the supremum of its power function

$$\beta(C; \delta_{ij}, \vartheta_{ij}) = P(T_{ave} > C | \delta_{ij}, \vartheta_{ij}) \tag{3.4}$$

given a prespecified critical point C and evaluated at $\vartheta_{ij} = 0$ for all $(i,j) \in \mathbb{G}$, over all possible values for δ_{ij} (Hung, Chi and Lipicky, 1993). As the supremum always occurs

at the extreme values of δ_{ij}, one can equivalently assume $|\delta_{ij}| = \infty$ for all $(i,j) \in \mathbb{G}$. This particularly offers a possibility to derive p-values for the AVE- and MAX-test from given observed values t_{ave} and t_{max} of the test statistics. They can be determined by the equations

$$p_{ave} = \int_0^\infty \Phi\left(-\sqrt{\frac{nAB}{1+\max\{A,B\}}}t_{ave}w\right) dQ(w) \tag{3.5}$$

$$p_{max} = 1 - \int_0^\infty \mathrm{E}\left[\Phi(\sqrt{n}t_{max}w + Z)^2\right] \mathrm{E}\left[\Phi(\sqrt{n}t_{max}w + Z)^4\right] dQ(w), \tag{3.6}$$

where Q denotes the c.d.f. of $\hat{\sigma}/\sigma$, Φ is the standard normal c.d.f. and Z a random variable with distribution according to Φ (Hung, Chi and Lipicky, 1993 and Hung, 1994). An estimate of the population variance σ^2 is needed in both (3.5) and (3.6); if σ^2 is known or the total sample size is sufficiently large, the integral corresponding to the distribution function Q can be omitted in the calculation of the p-values from the observed values of the test statistics.
In settings with $\delta_{ij} \approx 0$ for any $(i,j) \in \mathbb{G}$, the actual type I error does not match α exactly as (3.5) and (3.6) are approximate representations only. This is due to the assumption that $|\delta_{ij}| = \infty$ for all $(i,j) \in \mathbb{G}$. The resulting p-values are biased towards conservative test results, which turns the approximation out as a kind of "worst case" assumption assuring that the type I error is kept by the given level α.

From a practical point of view, it is for several reasons often desirable to allocate individuals to groups with *unequal* sample sizes: the trial might at the same time be intended to test for the efficacy of one or both of the component drugs, or some combinations might be tested at a greater power than others, e.g. for marketing reasons. The above approach to the global tests has therefore been extended to unbalanced bifactorial designs by Hung (2000). The approximation attained by the "worst case" assumption that $|\delta_{ij}| = \infty$ for all $(i,j) \in \mathbb{G}$ is still needed as the supremum in (3.4) will occur at the extreme values of the nuisance parameters also in the unbalanced case. The min-statistic is now replaced by the representation

$$X_{ij}(\pi_{ij}) := \pi_{ij}(\sqrt{\lambda_{1ij}}Z_{ij} - \sqrt{1-\lambda_{1ij}}Z_{i0}) + (1-\pi_{ij})(\sqrt{\lambda_{2ij}}Z_{ij} - \sqrt{1-\lambda_{2ij}}Z_{0j}), \tag{3.7}$$

where $\lambda_{1ij} = \frac{n_{i0}}{n_{ab}+n_{i0}}$, $\lambda_{2ij} = \frac{n_{0j}}{n_{ab}+n_{0j}}$ and $Z_{ij} = \frac{(X_{ij}-\mu_{ij})}{\sqrt{n_{ij}}\sigma}$. For the AVE-test, Hung (2000) showed that it is no loss of generality to determine the supremum in (3.4) by maximizing the function $P\left(X(\pi) > C | \pi_{ij}, \vartheta_{ij} = 0\right)$ instead, where $X(\pi) := (rs)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij}(\pi_{ij})$

and the supremum is taken over the set

$$\Omega := \{\pi = (\pi_{11}, ..., \pi_{rs}) | (\pi_{ij}, \pi_{ik}, \pi_{lj}, \pi_{lk}) \neq (1,0,0,1) \text{ or } (0,1,1,0) \text{ for } i < l \text{ and } j < k\}$$

of all possible extreme values of δ_{ij} except those that will never occur. For a given $\pi \in \Omega$, the variance of $X(\pi)$ is obtained as

$$\begin{aligned} \Delta(\pi) &= (rs)^{-2} \sum_{i=1}^{r} \sum_{j=1}^{s} (\pi_{ij}\lambda_{1ij} + (1-\pi_{ij})\lambda_{2ij}) \\ &\quad + \sum_{i=1}^{r} \left(\sum_{j=1}^{s} \pi_{ij} \sqrt{1-\lambda_{1ij}} \right)^2 + \sum_{i=1}^{r} \left(\sum_{j=1}^{s} (1-\pi_{ij}) \sqrt{\lambda_{1-2ij}} \right)^2 . \end{aligned} \tag{3.8}$$

Hence the approximate test statistic $X(\pi)$ can be studentized dividing it by $\Delta(\pi)$. This results in the asymptotically standard normal random variable

$$P\left(\frac{X(\pi)}{\sqrt{\Delta(\pi)}} > \frac{C}{\sqrt{\Delta(\pi)}} \middle| \pi_{ij}, \vartheta_{ij} = 0 \right) = 1 - \Phi\left(\frac{C}{\sqrt{\Delta(\pi)}} \right) \tag{3.9}$$

which has got its supremum at the same point $\pi \in \Omega$ where the variance given in (3.8) is maximized, following from monotonicity properties of Φ. Given an observed test statistic t_{ave}, the p-value can therefore be calculated by

$$p_{ave} = 1 - \Phi\left(\frac{t_{ave}}{\sqrt{\max_{\pi \in \Omega} \Delta(\pi)}} \right).$$

The restrictions of the approximation are the same as for the balanced case: the type I error is below the nominal level α, i.e. the test performs conservative if $\delta_{ij} \approx 0$ for any $(i,j) \in \mathbb{G}$.

The p-value for the MAX-test is determined by $p_{max} = \min\{\tilde{p}_{ij} | (i,j) \in \mathbb{G}\}$, where $\tilde{p}_{ij}$ are the adjusted p-values from the multiple testing procedure that is given in the next section.

3.2 Multiplicity-adjusted approach

The question *which* combination treatment groups have got significantly better responses than both components requires evaluation of one min-test for each combination drug. As this is leading to $A \cdot B$ tests, some considerations on multiplicity adjustment are necessary. For a treatment group $(i,j) \in \mathbb{G}$, the adjusted p-value $\tilde{p}_{ij}$ is the probability for at least one of the p-values in the analysis to be equal or smaller than p_{ij} under the complete null

hypothesis. In this case, where all statistics are of the same type and all are lower-tailed, this is equivalent to

$$\tilde{p}_{ij} = P_{H_0}\left(\max_{(i',j')\in\mathbb{G}} T_{i'j'}^{min} \geq t_{ij}^{min}\right) \tag{3.10}$$

according to the definiton in (2.5), i.e. the joined c.d.f. of $\{T_{ij}^{min}\}_{(i,j)\in\mathbb{G}}$ under the complete null hypothesis is needed which depends on the unknown nuisance parameters δ_{ij}. For monotonicity reasons, the familywise type I error is bounded by the significance level α if the supremum of the power function at a prespecified point C is taken over all δ_{ij} with $\vartheta_{ij} = 0$ for all $(i,j) \in \mathbb{G}$. It is again no loss of generality to maximize the power function of $X_{ij}(\pi_{ij})$ over the set Ω instead. The correlation matrix of $\{X_{ij}(\pi_{ij})\}_{(i,j)\in\mathbb{G}}$ is obtained as $\mathbf{R} = \{\varrho_{ij,lm}\}$, where

$$\varrho_{ij,lm} = \begin{cases} 1 & i = l, j = m \\ \pi_{ij}\pi_{im}\sqrt{(1-\lambda_{1ij})(1-\lambda_{1im})} & i = l, j \neq m \\ (1-\pi_{ij})(1-\pi_{im})\sqrt{(1-\lambda_{2ij})(1-\lambda_{2im})} & i \neq l, j = m \\ 0 & i \neq l, j \neq m \end{cases}$$

As the min-test is based on the two-sample *t*-statistic, the joined distribution can now be approximated by $\Phi_{AB}^{\mathbf{R}}$, the c.d.f. of the $(A\cdot B)$-variate normal distribution with covariance matrix $\mathbf{R}$ as above, concluding that the p-value is determined by $\tilde{p}_{ij} = 1 - \Phi_{AB}^{\mathbf{R}}(t_{ij}^{min}\mathbb{1}_{AB})$ from an observed value t_{ij}^{min} of the test statistic.

The approximation resulting from the assumption that $|\delta_{ij}| = \infty$ for all $(i,j) \in \mathbb{G}$ is used for all parametric settings, even if in particular the actual nuisance parameter is in an environment of zero. This assures protection of the nominal significance level α, but the actual type I error will in general not exactly match α except for large values of $|\delta_{ij}|$. For a single min-test in a balanced as well as in an unbalanced 1x1-design, Hung (2000) showed different power values to occur depending on the nuisance parameter δ_{11}. If in particular $\delta_{11} \gg 0$, the min-test reduces to the single t-statistic $T_{11}^{min} = T_{11}^{A}$, while $T_{11}^{min} = T_{11}^{B}$ analogously holds for $\delta_{11} \ll 0$. For normal data, the p-values are then uniformly distributed on the interval $[0,1]$. If on the other hand δ_{11} is close to 0, this approximation does not hold as there is a high probability for $T_{11}^{A} < T_{11}^{B}$ though in fact $\mu_{11} - \mu_{10} > \mu_{11} - \mu_{01}$ or vice versa, where the probabilities for both are approximately equal. The p-values p_{11}^{A} and p_{11}^{B} of the single t-statistics are then both rectangular-shaped and hence for the p-value $p_{11}^{min} = \max\{p_{11}^{A}, p_{11}^{B}\}$ of the min-test, greater values are more likely to occur than smaller ones. In this sense, the distribution of p_{11}^{min} for the combination group $(1,1) \in \mathbb{G}$ will be exactly rectangular only for $|\delta_{11}| = \infty$ and simulation

experiments will obtain a type I error of approximately α only in settings with large values of $|\delta_{11}|$. However, the nominal level α is kept well for all choices of δ_{11}, but the test performs conservative if δ_{11} is in fact in an environment of zero.

Buchheister and Lehmacher (2006) proposed multiple testing procedures based on the closed test principle (Chapter 2) as an alternative approach to the test problem (3.1). This requires construction of a hypotheses system that is closed under intersection: the single "marginal" hypotheses for group $(i,j) \in \mathbb{G}$ are denoted by $H_0^{ij,i0} : \mu_{ij} = \mu_{i0}$ and $H_0^{ij,0j} : \mu_{ij} = \mu_{0j}$. In detail, three different approaches are possible.
First, Buchheister and Lehmacher (2006) defined the global intersection hypothesis as the intersection of all local hypotheses expressed in terms of the marginal hypotheses:

$$H_0 = \bigcap_{(i,j)\in\mathbb{G}} H_0^{ij} = \bigcap_{(i,j)\in\mathbb{G}} (H_0^{ij,i0} \cup H_0^{ij,0j}), \tag{3.11}$$

which is equivalent to (3.3). The intersections of the local hypotheses for all groups $(i,j) \in \mathbb{G}$ are tested by a step-down procedure which will be discussed in Chapter 2 and can be shown to keep the level α. Now, each of the local hypotheses in the intersection (3.11) is a union of two hypotheses, whereas standard methods for closed test procedures are constructed for intersections of the hypotheses themselves. Using rules known from set theory, the global intersection hypothesis can be expressed by a union of intersection hypotheses for which a generalized version of the min-test applies. Buchheister (2001) illustrated this for closed testing procedures on 2x3- and 3x3-designs.
Second, a hypotheses system closed under intersection can be based on the marginal hypotheses $H_0^{ij,i0}$ and $H_0^{ij,0j}$ themselves, comprising $\sum_{g=1}^{2AB} \binom{2AB}{g}$ hypotheses, i.e. a substantially higher number than in the first approach (Buchheister and Lehmacher, 2006). However, using this method, the above transformations of hypotheses are no more needed, making the involved hypotheses easier to test. The global intersection hypothesis of this test can be expressed by intersections of the marginal hypotheses for both component drugs:

$$\tilde{H}_0 = \left(\bigcap_{(i,j)\in\mathbb{G}} H_0^{ij,i0} \right) \cap \left(\bigcap_{(i,j)\in\mathbb{G}} H_0^{ij,0j} \right)$$

A test on $\tilde{H}_0$ is **not** equivalent to the AVE- or MAX-test because $\tilde{H}_0$ differs from the global hypothesis in (3.3); nevertheless, this procedure does apply to test the local hypotheses in (3.1).
As a third approach, Buchheister and Lehmacher (2006) proposed a separate closed

hypotheses system for each collection of marginal hypotheses $H_0^{ij,i0}$ and $H_0^{ij,0j}$. The level α is kept by both procedures due to the closed test principle, but multiplicity adjustment is needed as the same procedure is applied *twice*. For instance, the simultaneous step-down procedures can be performed at level $\alpha/2$, implying adjustment according to the Bonferroni method (Chapter 2). The global intersection hypotheses are

$$\tilde{H}_0^A = \bigcap_{(i,j)\in\mathbb{G}} H_0^{ij,i0} \quad \text{and} \quad \tilde{H}_0^B = \bigcap_{(i,j)\in\mathbb{G}} H_0^{ij,0j}$$

and therefore distinct from (3.3). This approach is also appropriate for the local problem, but does not include a direct test of the global hypothesis.

3.3 Higher-dimensional factorial designs

In recent years, combinations of more than two compounds have often been applied if their respective efficacy mechanisms are distinct, especially in cancer therapies. The theory of this thesis applies to designs with arbitrarily high dimensions in an analogous way but might be practically relevant essentially for two or three compounds. Each combination treatment group can be represented as a k-tuple on the grid $\mathbb{G} = \mathbb{G}_0 \setminus \{(i_1, ..., i_k) | i_1 = 0 \vee ... \vee i_k = 0\}$, where

$$\mathbb{G}_0 = \{0, ..., D_1\} \times ... \times \{0, ..., D_k\} \subset \mathbb{N}^k,$$

i.e. there are now $(D_1 + 1)...(D_k + 1)$ treatment groups to allocate resources to which may limit the practicability to trials on frequent disease patterns as e.g. hypertension. For sake of simplicity, denote any point on the grid $\mathbb{G}$ by $\gamma := (i_1, ..., i_k)$ and the marginal treatment groups according to γ by the projections $\gamma^j := (i_1, ..., i_{j-1}, 0, i_{j+1}, ..., i_k)$. The question whether all component drugs give a contribution to the overall response of a combination treatment group $\gamma \in \mathbb{G}$ is represented by the test problem

$$H_0^\gamma : \bigvee_{j=1}^{k} (\mu_\gamma \leq \mu_{\gamma^j}) \quad \text{vs.} \quad H^\gamma : \bigwedge_{j=1}^{k} (\mu_\gamma > \mu_{\gamma^j}). \tag{3.12}$$

As a modification of the min-test that is reasonable for this local problem and a particular group $\gamma \in \mathbb{G}$, consider the test statistic $T_\gamma^{min} = \min_{j=1,...,k} T_\gamma^j$ with

$$T_\gamma^j = \frac{\bar{X}_\gamma - \bar{X}_{\gamma^j}}{\sqrt{\frac{\hat{\sigma}_\gamma^2}{n_\gamma} + \frac{\hat{\sigma}_{\gamma^j}^2}{n_{\gamma^j}}}}. \tag{3.13}$$

The c.d.f. of this will be shown to depend on the distance of the population means $\mu_{\gamma^1}, ..., \mu_{\gamma^k}$ in the marginal treatment groups as discussed in terms of the nuisance parameters δ_{ij} for the bifactorial case. In a general k-dimensional setting, the nuisance parameters can be denoted as a vector

$$\delta_\gamma = \begin{pmatrix} \mu_{\gamma^1} - \mu_{\gamma^2} \\ \vdots \\ \mu_{\gamma^1} - \mu_{\gamma^k} \\ \mu_{\gamma^2} - \mu_{\gamma^3} \\ \vdots \\ \mu_{\gamma^2} - \mu_{\gamma^k} \\ \vdots \\ \mu_{\gamma^{k-1}} - \mu_{\gamma^k} \end{pmatrix} \tag{3.14}$$

of length $\binom{k}{2}$ from which the nuisance parameters in the bifactorial design are obtained as a special case as $\mu_{\gamma^1} - \mu_{\gamma^2} = \mu_{i0} - \mu_{0j}$ for $\gamma = (i, j) \in \mathbb{G}$. On the other hand, the global test problem (3.3) can be generalized by

$$H_0 : \forall \gamma \in \mathbb{G} : \bigvee_{j=1}^{k} (\mu_\gamma \leq \mu_{\gamma^j}) \quad \text{vs.} \quad H_1 : \exists \gamma \in \mathbb{G} : \bigwedge_{j=1}^{k} (\mu_\gamma > \mu_{\gamma^j}), \tag{3.15}$$

which now reflects the question if *any* dose combination has got the desired property that all k component drugs give a contribution to the overall effect. Generalized forms of AVE- or MAX-statistics, that is to say

$$T_{ave} = (D_1 ... D_k)^{-1} \sum_{\gamma \in \mathbb{G}} T_\gamma^{min} \quad \text{and} \quad T_{max} = \max_{\gamma \in \mathbb{G}} T_\gamma^{min},$$

are possible approaches to test the global null hypothesis (3.15) for the k-factorial case. The methods proposed by Hung, Chi and Lipicky (1993), Hung (2000), Buchheister and Lehmacher (2006) do not cover this general approach for neither the global nor the multiple local test problems. In principle, the theory can be generalized to the $k \geq 3$ case which remains as an unsolved problem up to now. For the resampling methods discussed in the next chapter, this is not substantially more complicated than for $k = 2$.

In the bifactorial case which is most important for practice, an approach with less analytical effort is desirable. Resampling-based multiple testing generally offers an alternative especially for arbitrary distributional conditions, e.g. skewness and heteroscedasticity. These cases are not covered by the bifactorial design theory of Hung and others. However, the considerations in the next chapter will show that the resampling-based approach

cannot give a satisfactory solution to the dramatic loss of power in the min-test that occurs whenever the response parameters in single-compound dose groups are close for a particular combination. It will turn out later on that this is a general problem in the nature of the min-test. Nevertheless, the resampling-based approach is more flexible according to the distribution of the data and its application is therefore more convenient.

4 Bootstrap approach to k-factorial designs

Simulation results from Hung (2000) showed that the adjusted p-values from the methods of Chapter 1 tend to be very conservative in settings where no extreme nuisance parameters occur. In addition, the analytical solution proposed by Hung, Chi and Lipicky (1993) and Hung (2000) is applicable only in the special (but important) case where the underlying data are normally distributed with the variances assumed to be equal over the treatment groups. It is desirable to weaken the normality and especially the homoscedasticity assumption, using the actual empirical distribution of the data instead. For sake of simplification and clearness, a solution is needed that does not require cumbersome derivations of the power functions or the correlation matrix. Furthermore, no theory for tests on the generalized null hypotheses (3.12) and (3.15) with arbitrary dimensionality is available up to now, e.g. for the AVE- and MAX-test or multiple procedures for testing all treatment groups $\gamma \in \mathbb{G}$ where $k > 2$. This appears to be substantially more complicated than even for the bifactorial case discussed by Hung, Chi and Lipicky (1993) and Hung (2000). A resampling-based approach to the problem is therefore presented that is computationally intensive but has got the desired features.
To make resampling applicable to a wide range of multifactorial designs and to introduce a notation that is valid in a general sense, the dimensionality will not be specified from now on. A k-factorial design is considered where the treatment groups are represented as cells on the grid $\mathbb{G} = \{1, ..., D_1\} \times ... \times \{1, ..., D_k\} \subset \mathbb{N}^k$. The denotations γ for a combination group in $\mathbb{G}$ and γ^j, $j = 1, ..., k$, for specification of the marginal treatment groups are carried over from Section 3.3. In the introduction to resampling-based methods, permutation resampling has been shown to be more accurate because these methods are exact given the observed data and are equivalent to Fisher's exact test when applied to binary data. As the hypotheses of interest in connection with the mint-test have historically been stated in a parametric sense, the permutation methods do not apply to these. They are limited to hypotheses where two *distributions* are to be tested for equality, whereas the bootstrap is suitable for parametric hypotheses like (3.12). In the k-factorial design, the bootstrap procedure outlined by Algorithm 2.1 is therefore considered the best approach for this particular application, i.e. resampling will now be done by drawing *with replacement* samples from the data. This method has been shown

to be slightly more powerful when based on studentized statistics (Efron and Tibshirani, 1993). The bifactorial and trifactorial cases are obtained as special cases from this; evaluation of the proposed algorithms in terms of type I error and power will therefore be given for $k = 2$ and $k = 3$.

4.1 Bootstrapping the min-test

According to the first guideline for bootstrap testing in univariate situations reported in Section 2.5.1, resampling has to be done in a way that reflects the null hypothesis even if it is not satisfied by the population. As the distribution of the min-statistic under H_0^γ depends on the nuisance parameters δ_γ in the population, these have to be taken as a part of the null hypothesis; i.e. the resampling procedure must reflect

$$H_0^\gamma : (\mu_\gamma - \max\{\mu_{\gamma^1}, ..., \mu_{\gamma^k}\} = 0) \wedge \begin{pmatrix} \mu_{\gamma^1} - \mu_{\gamma^2} \\ \vdots \\ \mu_{\gamma^1} - \mu_{\gamma^k} \\ \mu_{\gamma^2} - \mu_{\gamma^3} \\ \vdots \\ \mu_{\gamma^2} - \mu_{\gamma^k} \\ \vdots \\ \mu_{\gamma^{k-1}} - \mu_{\gamma^k} \end{pmatrix} =: \delta_\gamma \tag{4.1}$$

as an extended version of the hypothesis in (3.12). The most natural approach to this is to estimate δ_γ by the sample mean differences and use this value for generating a resampled data set reflecting H_0^γ. Equivalently, the data of the marginal treatment groups can be resampled without centering by their mean; the data from the combination group γ are then centered by the maximum of the marginal group means to represent the first part of (4.1). The following algorithm performs the min-test for all combination groups $\gamma \in \mathbb{G}$ and involves adjustment for the multiple hypotheses problem as discussed in Section 2.5. This and the following algorithms will be given in a generalized form (left panel) as well as for the important $k = 2$ case (right panel).

Algorithm 4.1 *(Multiple min-tests with estimated nuisance parameters)*

(1)	*Initialize counting variables* $z_\gamma = 0$ *for* $\gamma \in \mathbb{G}$*. Calculate the min-statistics* T_γ^{min} *from the data* $\mathbf{X}_\gamma$ *and* $\{\mathbf{X}_{\gamma^j}\}_{j=1,...,k}$ *following equation* (3.13).	*Initialize counting variables* $z_{ij} = 0$ *for* $(i,j) \in \mathbb{G}$*. Calculate the estimates* $\hat{\delta}_{ij}$ *and the min-statistics* T_{ij}^{min} *from the data* $\mathbf{X}_{ij}$*,* $\mathbf{X}_{i0}$ *and* $\mathbf{X}_{0j}$ *following equation* (3.2).
(2)	*Generate* $(D_1+1)\cdot ...\cdot (D_k+1)$ *bootstrap samples* $\mathbf{X}_\gamma^*$ *that reflect the null hypothesis* (4.1)*: resample* $\mathbf{X}_{\gamma^j}^*$ *from* $\mathbf{X}_{\gamma^j}$ *for* $j = 1, ..., k$ *and* $\mathbf{X}_\gamma^*$ *from* $\mathbf{X}_\gamma - \bar{\mathbf{X}}_\gamma + \max\{\bar{\mathbf{X}}_{\gamma^1}, ..., \bar{\mathbf{X}}_{\gamma^k}\}$ *with replacement.*	*Generate* $(A+1)\cdot(B+1)$ *bootstrap samples* $\mathbf{X}_{ij}^*$ *that reflect* (4.1)*: resample* $\mathbf{X}_{ij}^*$ *from* $\mathbf{X}_{ij} - \bar{\mathbf{X}}_{ij} + \hat{\delta}_{ij}$ *with replacement. If* $\hat{\delta}_{ij} \geq 0$*, resample* $\mathbf{X}_{i0}^*$ *from* $\mathbf{X}_{i0} - \bar{\mathbf{X}}_{i0} + \hat{\delta}_{ij}$ *and* $\mathbf{X}_{0j}^*$ *from* $\mathbf{X}_{0j} - \bar{\mathbf{X}}_{0j}$*; if* $\hat{\delta}_{ij} < 0$*, resample* $\mathbf{X}_{i0}^*$ *from* $\mathbf{X}_{i0} - \bar{\mathbf{X}}_{i0}$ *and* $\mathbf{X}_{0j}^*$ *from* $\mathbf{X}_{0j} - \bar{\mathbf{X}}_{0j} + \hat{\delta}_{ij}$*.*
(3)	*Calculate the min-test statistics* T_γ^{min*} *from the resampled vectors* $\mathbf{X}_\gamma^*$ *and* $\{\mathbf{X}_{\gamma^j}^*\}_{j=1,...,k}$ *following* (3.13)*. Check whether* $\max_{\gamma' \in \mathbb{G}} T_{\gamma'}^{min*} \geq T_\gamma^{min}$ *and in case increase* z_γ *by 1.*	*Calculate the min-test statistics* T_{ij}^{min*} *from the resampled vectors* $\mathbf{X}_{ij}^*$*,* $\mathbf{X}_{i0}^*$ *and* $\mathbf{X}_{0j}^*$ *following* (3.2)*. Check whether* $\max_{(i',j') \in \mathbb{G}} T_{i'j'}^{min*} \geq T_{ij}^{min}$ *and in case increase* z_{ij} *by 1.*
(4)	*Repeat steps (2) and (3)* N *times and estimate the adjusted* p*-values by* $\hat{\tilde{p}}_\gamma^{(N)} = \frac{z_\gamma}{N}$*.*	*Repeat steps (2) and (3)* N *times and estimate the adjusted* p*-values by* $\hat{\tilde{p}}_{ij}^{(N)} = \frac{z_{ij}}{N}$*.*

The correlation structure from the data is implicitly used in a way that statistics with non-vanishing correlation are resampled from overlapping data sets. Note that it is no loss of generality to resample under the complete null hypothesis $\mathbf{H}_0^{\mathbb{G}} = \cap_{\gamma \in \mathbb{G}} H_0^\gamma$ if the test statistics satisfy the subset pivotality condition (Westfall and Young, 1993). If $K \subset \mathbb{G}$ is a subset of the design grid where all H_0^κ are true for $\kappa \in K$, the joint distribution of the test statistics $\{T_\kappa\}_{\kappa \in K}$ depends on the sample size allocation, the correlation matrix of the statistics and the nuisance parameters $\{\delta_\kappa\}_{\kappa \in K}$. In particular, the distribution does not depend on which particular subset of null hypotheses is true, i.e. on the values of the remaining means $\{\mu_\kappa\}_{\kappa \notin K}$. Hence the subset pivotality condition is satisfied for testing multiple hypotheses by the min-test.

To get an idea of the performance of bootstrap-based min-tests, it is convenient to consider a design with $k = 2$ or $k = 3$ and only one combination group and its respective components. This does obviously not require any multiplicity adjustment. Denote the

Algorithm	n	δ=0.0	δ=0.1	δ=0.2	δ=0.3	δ=0.5	δ=0.8
4.1	10	0.0343	0.0419	0.0517	0.0551	0.0651	0.0706
	25	0.0296	0.0465	0.0589	0.0684	0.0684	0.0566
	50	0.0334	0.0532	0.0660	0.0725	0.0583	0.0501
	100	0.0340	0.0615	0.0698	0.0630	0.0551	0.0525
	250	0.0319	0.0678	0.0639	0.0556	0.0488	0.0488
4.1 (δ known)	10	0.0520	0.0504	0.0506	0.0504	0.0495	0.0482
	25	0.0496	0.0506	0.0524	0.0522	0.0496	0.0486
	50	0.0507	0.0487	0.0491	0.0478	0.0505	0.0520
	100	0.0505	0.0494	0.0495	0.0530	0.0510	0.0523
	250	0.0510	0.0488	0.0497	0.0526	0.0503	0.0514
4.2	10	0.0140	0.0180	0.0211	0.0266	0.0350	0.0458
	25	0.0130	0.0200	0.0269	0.0379	0.0497	0.0473
	50	0.0117	0.0220	0.0350	0.0406	0.0476	0.0521
	100	0.0133	0.0284	0.0410	0.0483	0.0508	0.0491
	250	0.0120	0.0343	0.0471	0.0489	0.0493	0.0510

Table 4.1: *Results of* $25,000$ *simulations of the* 0.05 *level min-test for a 2x2 bifactorial design using Algorithms 4.1 and 4.2 with* $N = 15,000$ *bootstrap iterations each. For Algorithm 4.1, the nominal significance level is exceeded for nuisance parameters* $\delta > 0$*. Algorithm 4.2 performs very conservative for small values of* δ*. The second panel shows results from evaluation of Algorithm 4.1, but with a known value of the nuisance parameter. The actual type I error is then close to the nominal level* α *for all choices of* δ*.*

population means of the treatment groups involved by μ_{11}, μ_{10} and μ_{01} with $\delta = \mu_{10} - \mu_{01}$ in the $k = 2$ case. For simulation under the null hypothesis, data are then sampled from normal populations with variance $\sigma^2 = 1$ and means $\mu_{10} = 0$ and $\mu_{01} = \mu_{11} = \delta$. For $k = 3$, denote the means by μ_{111}, μ_{110}, μ_{101} and μ_{011}, where the data are simulated from populations with several settings for the marginal means μ_{110}, μ_{101} and μ_{011} according to the columns of Table 4.2 and $\mu_{111} = \max\{\mu_{110}, \mu_{101}, \mu_{011}\}$ reflecting the first part of (4.1). The results of the studies on these examples are summarized in Table 4.1 for $k = 2$ and in Table 4.2 for $k = 3$, using several algorithms. Table 4.1 shows that the proposed bootstrap approach to the min-test keeps the given level $\alpha = 0.05$ well for a nuisance parameter $\delta = 0$, but performs *anticonservative* for settings where $\delta > 0$. The same is observed for $k = 3$, where α is protected well if all marginal means are equal. The three components of the nuisance parameter vector δ are the same but in permuted order for the third and forth column. However, denoting the ordered values of the marginal means by $\mu_{(1)}$, $\mu_{(2)}$ and $\mu_{(3)}$, the parameter $|\mu_{(3)} - \mu_{(2)}|$ in the example has got the two distinct values 0.1 and 0.4 and the simulation results are also distinct in these settings. Furthermore, for the case where $|\mu_{(3)} - \mu_{(2)}| = 0.1$, the type I error additionally depends on the

Algorithm	n	(0.0,0.0,0.0)	(0.2,0.2,0.2)	(0.5,0.4,0.0)	(0.5,0.1,0.0)	(0.5,0.4,0.4)	(0.5,0.1,0.1)
4.1	10	0.0241	0.0244	0.0474	0.0694	0.0362	0.0699
	25	0.0244	0.0276	0.0589	0.0851	0.0426	0.0806
	50	0.0269	0.0283	0.0564	0.0783	0.0509	0.0769
	100	0.0242	0.0310	0.0647	0.0577	0.0597	0.0664
	250	0.0222	0.0246	0.0781	0.0483	0.0763	0.0516
4.2	10	0.0049	0.0049	0.0140	0.0276	0.0082	0.0224
	25	0.0060	0.0050	0.0189	0.0383	0.0110	0.0340
	50	0.0053	0.0059	0.0224	0.0476	0.0131	0.0456
	100	0.0040	0.0060	0.0277	0.0456	0.0178	0.0506
	250	0.0048	0.0055	0.0349	0.0519	0.0247	0.0522

Table 4.2: *Results of* $25,000$ *simulations for the* 0.05 *level min-test on a 2x2x2 trifactorial design using Algorithm 4.1 and 4.2 with* $N = 15,000$ *bootstrap iterations each. The actual type I error is given for various combinations of the marginal means and depends primarily on the parameter* $|\mu_{(3)} - \mu_{(2)}|$*: for Algorithm 4.1, the nominal level is not protected if* $|\mu_{(3)} - \mu_{(2)}| > 0$*. Applying Algorithm 4.2, the test is very conservative if* $|\mu_{(3)} - \mu_{(2)}|$ *is in an environment of zero.*

difference $|\mu_{(3)} - \mu_{(1)}|$, whereas the latter has no impact for $|\mu_{(3)} - \mu_{(2)}| = 0.4$ (last two columns). It can therefore be supposed that in general, the components of the nuisance parameter vector δ are of unequal importance for the test level; more exactly, the actual type I error primarily depends on the parameter $|\mu_{(k)} - \mu_{(k-1)}|$ and the influence of the parameter $|\mu_{(k)} - \mu_{(k-2)}|$ is getting stronger as $|\mu_{(k)} - \mu_{(k-1)}|$ decreases.
The simulation results in Tables 4.1 and 4.2 exemplarily show for $k = 2$ and $k = 3$ that the min-test based on Algorithm 4.1 exceeds the given significance level if δ is estimated from the data. It will be shown later on that this stems from the poor accuracy of the estimate for δ. This is confirmed by an additional simulation that has been performed for $k = 2$, using a prespecified value of δ to resample under the *true* null hypothesis. The nuisance parameter δ is assumed to be *known* and the data vectors are centered according to δ prior to the resampling procedure in Algorithm 4.1 such that hypothesis (4.1) is reflected. These results are additionally given in Table 4.1: the type I error of the test approximately matches the level α for all values of δ and n but is of no use for practical purposes as knowledge of δ cannot be assumed.

As the p-values might be *too small* for some settings of the nuisance parameters, the bootstrap approach to the min-test proposed in Algorithm 4.1 is **not** suitable to increase the power. From a regulatoric point of view, a better way to protect the significance level is to involve the assumption that, as proposed by Hung (1993) for $k = 2$, $|\mu_{(k)} - \mu_{(k-1)}| = \infty$

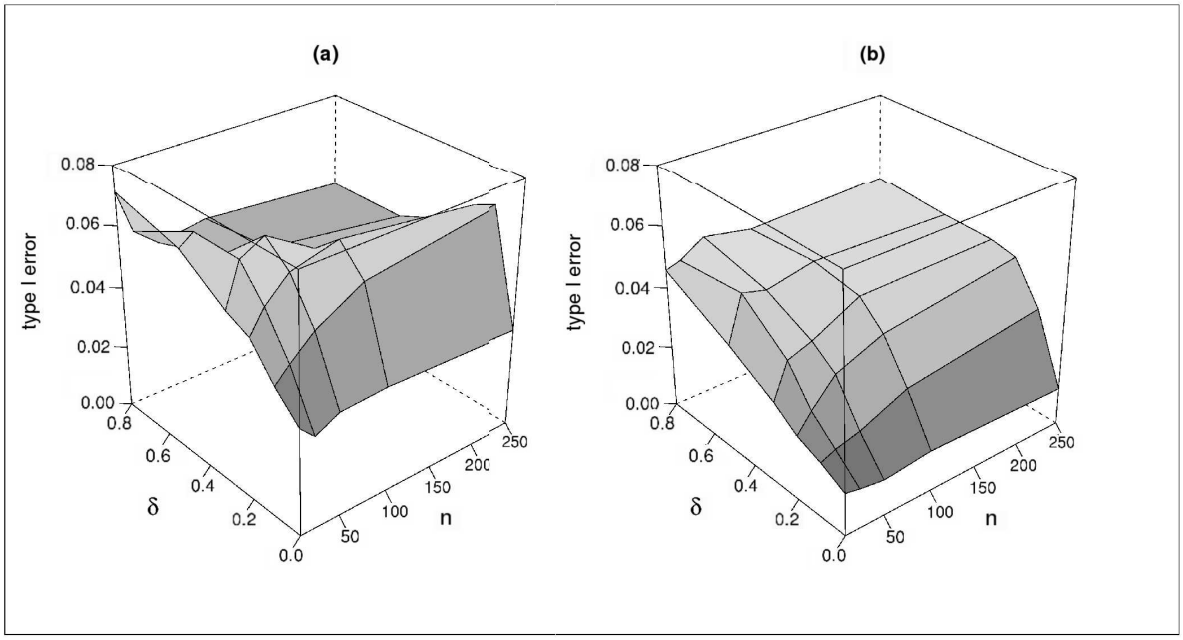

Figure 4.1: *Vizualisation of the results in Table 4.1.* ***(a)*** *Simulations are based on Algorithm 4.1. The nominal level $\alpha = 0.05$ is exceeded for nuisance parameters $\delta > 0$.* ***(b)*** *Evaluation of Algorithm 4.2: the type I error is much smaller than $\alpha = 0.05$, particularly if δ is close to zero. Small sample sizes yield more conservative tests than large-sample designs.*

for all $\gamma \in \mathbb{G}$ even if any of these parameters is in fact in an environment of 0. In practice, this can be achieved by simulating the distribution of T_γ^{min} using single two-sample t-statistics based on two groups each. The combination group is compared to the marginal group with the largest sample mean, i.e. the resampling-based approach is based on $\mathbf{X}_\gamma - \bar{\mathbf{X}}_\gamma$ and $\mathbf{X}_{\gamma^j} - \bar{\mathbf{X}}_{\gamma^j}$ if $\bar{\mathbf{X}}_{\gamma^j} = \max\{\bar{\mathbf{X}}_{\gamma^1}, ..., \bar{\mathbf{X}}_{\gamma^k}\}$. For the $k = 2$ case, the marginal group with the larger sample mean can equivalently be determined according to the *signs* of the $\hat{\delta}_{ij}$, i.e. the calculations are based on $\mathbf{X}_{ij} - \bar{\mathbf{X}}_{ij}$ and $\mathbf{X}_{i0} - \bar{\mathbf{X}}_{i0}$ if $\hat{\delta}_{ij} \geq 0$ and on $\mathbf{X}_{ij} - \bar{\mathbf{X}}_{ij}$ and $\mathbf{X}_{0j} - \bar{\mathbf{X}}_{0j}$ if $\hat{\delta}_{ij} < 0$. The algorithm is given as a modified version of Algorithm 4.1, again in a general form for multiple inferences on a factorial grid $\mathbb{G}$ as well as for $k = 2$:

Algorithm 4.2 *(Multiple min-tests with conservative assumption)*

(1)	*Initialize counting variables* $z_\gamma = 0$ *for* $\gamma \in \mathbb{G}$*. Determine the indices* j *with* $\bar{\mathbf{X}}_{\gamma^j} = \max\{\bar{\mathbf{X}}_{\gamma^1}, ..., \bar{\mathbf{X}}_{\gamma^k}\}$ *and* $T_\gamma^{min} = T_\gamma^j$.	*Initialize counting variables* $z_{ij} = 0$ *for* $(i,j) \in \mathbb{G}$ *and calculate the statistics* $T_{ij}^{min} = \min\{T_{i0}, T_{0j}\}$.
(2)	*Generate* $(D_1+1)\cdot ... \cdot (D_k+1)$ *bootstrap samples* $\mathbf{X}_\gamma^*$*: resample* $\mathbf{X}_\gamma^*$ *from* $\mathbf{X}_\gamma - \bar{\mathbf{X}}_\gamma + \max\{\bar{\mathbf{X}}_{\gamma^1}, ..., \bar{\mathbf{X}}_{\gamma^k}\}$ *and* $\mathbf{X}_{\gamma^j}^*$ *from* $\mathbf{X}_{\gamma^j}$.	*Generate* $(A+1)\cdot(B+1)$ *bootstrap samples: resample* $\mathbf{X}_{ij}^*$ *from* $\mathbf{X}_{ij} - \bar{\mathbf{X}}_{ij}$*. If* $\hat{\delta}_{ij} \geq 0$*, resample* $\mathbf{X}_{i0}^*$ *from* $\mathbf{X}_{i0} - \bar{\mathbf{X}}_{i0}$*, otherwise* $\mathbf{X}_{0j}^*$ *from* $\mathbf{X}_{0j} - \bar{\mathbf{X}}_{0j}$.
(3)	*Determine the min-test statistics* $T_\gamma^{min*} = T_{\gamma^j}$*. Check whether* $\max_{\gamma' \in \mathbb{G}} T_{\gamma'}^{min*} \geq T_\gamma^{min}$ *and in case increase* z_γ *by 1.*	*Determine the min-test statistics* T_{ij}^{min*}*. Check whether* $\max_{(i',j') \in \mathbb{G}} T_{i'j'}^{min*} \geq T_{ij}^{min}$ *and in case increase* z_{ij} *by 1.*
(4)	*Repeat steps (2) and (3)* N *times. The estimated adjusted* p*-values are then* $\hat{\tilde{p}}_\gamma^{(N)} = \frac{z_\gamma}{N}$.	*Repeat steps (2) and (3)* N *times. The estimated adjusted* p*-values are then* $\hat{\tilde{p}}_{ij}^{(N)} = \frac{z_{ij}}{N}$.

This method has been evaluated by the same simulation experiments as performed for Algorithm 4.1. The results are given in the last panels of Table 4.1 and Table 4.2 for $k = 2$ and $k = 3$, respectively: in the $k = 2$ case, the p-values estimated by Algorithm 4.2 tend to be conservative if δ is in an environment of zero, whereas the type I error is close to α if δ is large. These results essentially agree to those obtained by Hung (2000) from the evaluation of his analytical min-test approach. Recalling the evaluation of Algorithm 4.1, it is no surprise that for $k = 3$, the type I error of the min test primarily depends on how well the largest marginal means can be distinguished. If, in particular, the parameter $|\mu_{(3)} - \mu_{(2)}|$ is close to zero, the test performs much more conservative than for large values of $|\mu_{(3)} - \mu_{(2)}|$. In addition, the impact of the parameter $|\mu_{(3)} - \mu_{(1)}|$ is stronger if $|\mu_{(3)} - \mu_{(2)}|$ is small. It can be supposed that this holds for the parameters $|\mu_{(k)} - \mu_{(k-1)}|$ and $|\mu_{(k)} - \mu_{(k-2)}|$ in a design with general choice of k.

There are a couple of possible modifications for Algorithm 4.2 which are essentially equivalent but make different use of the data. One of these is to calculate the test statistics alternately from both marginal groups, reflecting the fact that in case of small values of the nuisance parameters, it is unknown which of the two marginal groups has in fact larger mean. This approach avoids the problem of throwing away the data from the

		$\delta = 0.0$		$\delta = 0.1$		$\delta = 0.3$		$\delta = 0.5$	
Population	ϑ	Hung	Bootstrap	Hung	Bootstrap	Hung	Bootstrap	Hung	Bootstrap
Normal, homoscedastic	0.0	0.0116	0.0117	0.0223	0.0220	0.0425	0.0406	0.0506	0.0476
	0.2	0.1256	0.1278	0.1776	0.1830	0.2475	0.2484	0.2546	0.2626
	0.3	0.2737	0.2735	0.3572	0.3506	0.4337	0.4256	0.4396	0.4361
	0.4	0.4728	0.4757	0.5713	0.5599	0.6214	0.6228	0.6382	0.6283
	0.5	0.6851	0.6907	0.7569	0.7482	0.7976	0.7996	0.8001	0.8002
	0.6	0.8505	0.8443	0.8934	0.8857	0.9074	0.9051	0.9079	0.9059
Normal, heteroscedastic	0.0	0.0118	0.0129	0.0299	0.0222	0.0617	0.0465	0.0813	0.0487
	0.2	0.1272	0.1317	0.2022	0.1859	0.2666	0.2216	0.2908	0.2280
	0.3	0.2758	0.2806	0.3701	0.3493	0.4426	0.3897	0.4588	0.3799
	0.4	0.4631	0.4576	0.5518	0.5334	0.6025	0.5538	0.6200	0.5370
	0.5	0.6562	0.6419	0.7282	0.7023	0.7561	0.7184	0.7695	0.6996
	0.6	0.8034	0.7963	0.8550	0.8388	0.8698	0.8404	0.8742	0.8321
Lognormal, homoscedastic	0.0	0.0115	0.0176	0.0210	0.0291	0.0328	0.0567	0.0346	0.0672
	0.2	0.1447	0.1422	0.1950	0.2077	0.2591	0.2799	0.2595	0.3107
	0.3	0.3097	0.2936	0.3963	0.3682	0.4640	0.4627	0.4677	0.4892
	0.4	0.5354	0.4790	0.6069	0.5460	0.6703	0.6423	0.6753	0.6520
	0.5	0.7337	0.6419	0.7924	0.7001	0.8241	0.7604	0.8377	0.7785
	0.6	0.8620	0.7706	0.8994	0.8124	0.9218	0.8585	0.9277	0.8659
Lognormal, heteroscedastic	0.0	0.0115	0.0151	0.0208	0.0328	0.0424	0.0605	0.0510	0.0667
	0.2	0.1240	0.1445	0.1864	0.2076	0.2592	0.2923	0.2940	0.2913
	0.3	0.2884	0.3063	0.3693	0.3879	0.4536	0.4600	0.4857	0.4511
	0.4	0.5071	0.5195	0.6022	0.5914	0.6685	0.6356	0.6984	0.6284
	0.5	0.7143	0.7224	0.7963	0.7661	0.8358	0.7947	0.8526	0.7729
	0.6	0.8773	0.8568	0.9247	0.8855	0.9396	0.8879	0.9380	0.8693

Table 4.3: *Results of* $10,000$ *simulations for the power of the* 0.05 *level min-test in a 2x2 bifactorial design using Algorithm 4.2 with* $n = 50$ *and* $N = 15,000$ *bootstrap iterations each. The power is monotonically increasing in* δ *also for non-vanishing values of the primary parameter* ϑ*.*

wrong group that might occur if, for instance, $\hat{\delta} > 0$ though in fact $\delta < 0$. A similar approach is to choose the bootstrap samples from the pooled data of both marginal groups. However, as the resulting bootstrap distribution will approximately be a t-distribution for both methods, the simulation results under the null hypothesis (Table 4.1) as well as under the alternative and for different distributional settings (Table 4.3) are very similar. These results are therefore not shown.

For bifactorial designs with just one combination group and a single min-test, simulations on Algorithm 4.2 have also been performed under the alternative hypothesis for $n = 50$. Evidently, a greater power is expected for larger sample sizes in any case, but the analysis is focused on the influence of the primary parameter $\vartheta = \mu_{11} - \max\{\mu_{10}, \mu_{01}\}$ and the nuisance parameter δ. As mentioned in Chapter 2, the bootstrap allows for skewed

and heteroscedastic data. To compare the performance to that of the analytical approach of Hung (2000) which involves the multivariate normal distribution, the results are given for both methods. To attain that the pooled variance over the three groups $(1,1)$, $(1,0)$ and $(0,1)$ is still 1, the variances on which the simulations of heteroscedastic data are based must be defined by

$$\begin{aligned}\sigma_{10} &= a,\\ \sigma_{01} &= a(1+\delta)\\ \sigma_{11} &= a(1+\delta+\vartheta),\end{aligned}$$

where $a = \sqrt{3/\left(1+(1+\delta)^2+(1+\delta+\vartheta)^2\right)}$. The simulation results under the alternative have been summarized in Table 4.3: as expected, the power primarily depends on ϑ, but for fixed values of ϑ, it is also monotonically increasing in δ. The power is sufficient for large primary parameters even if δ is in an environment of zero. For the normal and homoscedastic case, there is no definite distinction between the analytical approach and the bootstrap. When the equal variances assumption must be dropped, the analytical method obviously exceeds the nominal level α under the null hypothesis if the nuisance parameters are large. This problem is not present when using Algorithm 4.2 where the power is slightly smaller in all parametric settings for normal data. If the underlying population is lognormal, the power for small effect sizes is larger when using Algorithm 4.2 for the homoscedastic as well as the heteroscedastic case. This includes the problem that under the null hypothesis, the significance level might be slightly exceeded by the bootstrap if the nuisance parameters are large. In contrast, the power of the analytical approach is better than that of the bootstrap for large effect sizes.

For sample size planning, the power of the tests on normal and homoscedastic data has been evaluated for several values of n where the samples in the respective groups are assumed to be equally-sized. The results for this are given in Table 4.4. As $\alpha = 0.05$ and $1-\beta = 0.8$ are usual choices for the error bounds, the sample sizes where the power is at least 0.8 should be chosen. In practice, the parameters are defined as a clinically relevant effect expressed in units of the supposed population standard deviations.

4.2 Resampling-based tests of global hypotheses

The remarks on the bootstrap approach for the min-test also apply to the AVE- and MAX-test on the global problem (3.15) that was previously discussed by Hung, Chi and Lipicky

	$\vartheta = 0.4$			$\vartheta = 0.6$		
Sample size	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0.0$	$\delta = 0.2$	$\delta = 0.4$
$n = 5$	0.2546	0.2391	0.2270	0.3454	0.3322	0.3219
$n = 15$	0.3960	0.3776	0.3630	0.6096	0.5903	0.5492
$n = 25$	0.5146	0.5079	0.4656	0.7704	0.7377	0.7139
$n = 50$	0.7375	0.6957	0.6511	0.9440	0.9217	0.9153
$n = 75$	0.8476	0.8187	0.7881	0.9873	0.9807	0.9765

Table 4.4: *Power simulations of the* 0.05 *level min-test for the normal and homoscedastic case and various effect parameters* ϑ *and* δ *for purpose of sample size planning. Assuming* $1 - \beta = 0.8$*, the samples sizes should be chosen from the table such that this power can be achieved by the test. The calculations are based on* $10,000$ *simulations using Algorithm 4.2 with* $N = 15,000$ *bootstrap iterations.*

(1993), Hung (2000), Hellmich and Lehmacher (2005), Buchheister and Lehmacher (2006). The powerful theory behind this was summarized in Chapter 3, but is restricted to $k = 2$ and certain distributional cases. Furthermore, it might not be feasible for practical purposes to handle with the rather technical analytical approach. In the following, the bootstrap methods are therefore extended to the AVE- and MAX-test, yielding an approach that needs no special theory on the power functions.
For the AVE-test, the null distribution of the random variable T_{ave} can be simulated by resampling from the original data. As the AVE-statistic is a mean of $D_1 \cdot ... \cdot D_k$ test statistics, it needs to be studentized dividing by the variance of T_{ave}. This has got the representation

$$\Delta = \frac{\left(\sum_{\gamma \in \mathbb{G}} (T_\gamma^{min})^2 - \frac{\left(\sum_{\gamma \in \mathbb{G}} T_\gamma^{min} \right)^2}{D_1 \cdot ... \cdot D_k} \right)}{(D_1 \cdot ... \cdot D_k)((D_1 \cdot ... \cdot D_k) - 1)} \tag{4.2}$$

for the given sample as well as for each bootstrapped data set generated throughout the simulation, avoiding calculation of the variance from (3.8). The p-value is then the probability

$$p_{ave} = P_{H_0} \left(\frac{T_{ave}}{\sqrt{\Delta}} \geq \frac{t_{ave}}{\sqrt{\hat{\Delta}}} \right),$$

where $\hat{\Delta}$ denotes the estimated value of Δ from the observed values t_γ^{min} of the min-statistics. According to the nuisance parameters δ_γ, the same problems obviously arise for the AVE-test as for the single min-tests and the multiple testing procedures in the previous sections. Keeping in mind that the min-test fails to protect the nominal significance level if $|\mu_{(k)} - \mu_{(k-1)}|$ is estimated from the data, the assumption $|\mu_{(k)} - \mu_{(k-1)}| = \infty$ for all $\gamma \in \mathbb{G}$ will be applied again and conservative results are expected, regarding the

evaluation of Algorithm 4.2. In detail, the calculations for a group $\gamma \in \mathbb{G}$ are based on $\mathbf{X}_\gamma$ and $\mathbf{X}_{\gamma^j}$ if $\bar{\mathbf{X}}_{\gamma^j} = \max\{\bar{\mathbf{X}}_{\gamma^1}, ..., \bar{\mathbf{X}}_{\gamma^k}\}$. The following bootstrap-based algorithm computes the p-value for the AVE-test, avoiding the derivation of the c.d.f. of T_{ave}:

Algorithm 4.3 *(Resampling-based AVE-test)*

(1)	*Initialize a counting variable* $z = 0$. *For all* $\gamma \in \mathbb{G}$*, determine the index* j *with* $\bar{\mathbf{X}}_{\gamma^j} = \max\{\bar{\mathbf{X}}_{\gamma^1}, ..., \bar{\mathbf{X}}_{\gamma^k}\}$ *and define* $T_\gamma^{min} = T_\gamma^j$. *Calculate the variance* $\hat{\Delta}$ *and the AVE-statistic* $T_{ave} = (D_1 \cdot ... \cdot D_k)^{-1} \sum_{\gamma \in \mathbb{G}} T_\gamma^{min}$.	*Initialize a counting variable* $z = 0$. *Calculate* $T_{ij}^{min} = \min\{T_{i0}, T_{0j}\}$ *for all* $(i, j) \in \mathbb{G}$ *and the variance* $\hat{\Delta}$ *and* $T_{ave} = (AB)^{-1} \sum_{(i,j) \in \mathbb{G}} T_{(i,j)}^{min}$.
(2)	*Generate* $2D_1 \cdot ... \cdot D_k$ *bootstrap samples: resample* $\mathbf{X}_\gamma^*$ *from* $\mathbf{X}_\gamma - \bar{\mathbf{X}}_\gamma$ *and* $\mathbf{X}_{\gamma^j}^*$ *from* $\mathbf{X}_{\gamma^j} - \bar{\mathbf{X}}_{\gamma^j}$.	*Generate* $2AB$ *bootstrap samples: resample* $\mathbf{X}_{ij}^*$ *from* $\mathbf{X}_{ij} - \bar{\mathbf{X}}_{ij}$. *If* $\hat{\delta}_{ij} \geq 0$*, resample* $\mathbf{X}_{i0}^*$ *from* $\mathbf{X}_{i0} - \bar{\mathbf{X}}_{i0}$*, otherwise* $\mathbf{X}_{0j}^*$ *from* $\mathbf{X}_{0j} - \bar{\mathbf{X}}_{0j}$.
(3)	*Determine* $T_\gamma^{min*} = T_\gamma^{j*}$. *Calculate* $T_{ave}^* = (D_1 \cdot ... \cdot D_k)^{-1} \sum_{\gamma \in \mathbb{G}} T_\gamma^{min*}$ *and the sample variance* $\hat{\Delta}^*$ *from* (4.2). *Check whether* $(T_{ave}^*/\sqrt{\hat{\Delta}^*}) \geq (T_{ave}/\sqrt{\hat{\Delta}})$ *and in case increase* z *by 1.*	*Determine the min-statistics* T_{ij}^{min*}*, calculate* $T_{ave}^* = (AB)^{-1} \sum_{(i,j) \in \mathbb{G}} T_{ij}^{min*}$ *and the sample variance* $\hat{\Delta}^*$ *from* (4.2). *Check whether* $(T_{ave}^*/\sqrt{\hat{\Delta}^*}) \geq (T_{ave}/\sqrt{\hat{\Delta}})$ *and in case increase* z *by 1.*
(4)	*Repeat steps (2) and (3)* N *times. The estimate for the* p*-value is* $\hat{p}_{ave}^{(N)} = \frac{z}{N}$.	*Repeat steps (2) and (3)* N *times. The estimate for the* p*-value is* $\hat{p}_{ave}^{(N)} = \frac{z}{N}$.

The p-value for the MAX-test is obtained as the smallest adjusted p-value resulting from Algorithm 4.2 and therefore requires no further theory on numerical calculation.

Statistical power of Algorithm 4.3 and of the MAX-test based on Algorithm 4.2 have been evaluated by simulation studies. Following Hung (2000), the matrices E_1 and E_2 are involved to represent the effect sizes occuring under the alternative hypothesis in a simulated 4x3 bifactorial trial with two drugs evaluated in the doses $0, 1, 2, 3$ (drug A) and $0, 1, 2$ (drug B), respectively: let

$$E_1 := \begin{pmatrix} 0.0 & 0.1 & 0.3 & 0.6 \\ 0.2 & 0.5 & 0.6 & 0.9 \\ 0.5 & 0.8 & 0.8 & 0.9 \end{pmatrix} \quad \text{and} \quad E_2 := \begin{pmatrix} 0.0 & 0.1 & 0.3 & 0.6 \\ 0.2 & 0.25 & 0.7 & 1.0 \\ 0.5 & 0.65 & 0.9 & 1.0 \end{pmatrix},$$

Population	Effect	Variance	AVE-test		MAX-test	
			Hung (2000)	Bootstrap	Hung (2000)	Bootstrap
Normal	E_0	V_{hom}	0.0128	0.0537	0.0283	0.0245
	E_1	V_{hom}	0.7625	0.7920	0.5803	0.5170
		V_1	0.7731	0.7231	0.6518	0.5591
	E_2	V_{hom}	0.7719	0.7246	0.7099	0.6314
		V_2	0.7374	0.7054	0.6990	0.5831
Lognormal	E_0	V_{hom}	0.0117	0.0661	0.0198	0.0196
	E_1	V_{hom}	0.7250	0.6924	0.6040	0.4685
		V_1	0.7782	0.7050	0.6249	0.5414
	E_2	V_{hom}	0.7232	0.7740	0.7199	0.5544
		V_2	0.7491	0.7999	0.7074	0.5994

Table 4.5: *Simulation results for the* 0.05 *level AVE- and MAX-tests for several distributional cases. Under the null hypothesis,* $25,000$ *simulations were performed, whereas* $10,000$ *were considered enough under the alternative. The calculations are based on Algorithm 4.3 for the AVE-test and Algorithm 4.2 for the MAX-test with* $N = 15,000$ *bootstrap iterations. For comparison, the multivariate normal approach proposed by Hung (2000) was evaluated, where the implementation was overtaken from Hellmich and Lehmacher (2005).*

where E_1 represents the case that the effects have got a common value $\vartheta_{ij} = \mu_{ij} - \max\{\mu_{i0}, \mu_{0j}\} = 0.3$ for all $(i, j) \in \mathbb{G}$, whereas for E_2, the groups with $A = 1$ have smaller effects than the remaining combinations but still with a mean effect of 0.3 over the combination groups. Sample size allocation matrices as chosen by Hung (2000) will not be considered here as the focus of the simulations will primarily be on variations of the distributional shape of the data. The AVE- and MAX-test have been evaluated for normal versus lognormal populations and for homoscedastic as well as for heteroscedastic data. The equal variances case is represented by the matrix

$$V_{hom} := \begin{pmatrix} 1.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 \end{pmatrix},$$

whereas for the heteroscedastic case, the coefficient of variation is held constant over the treatment groups, i.e. the standard deviations linearly increase in the population means according to, say, $\sigma = 0.5 + 0.9\mu$. For both effect size matrices, the variances in the respective treatment groups can then be represented by one matrix each, that is to say

$$V_1 := \begin{pmatrix} 0.50 & 0.60 & 0.77 & 1.04 \\ 0.68 & 0.95 & 1.04 & 1.31 \\ 0.95 & 1.22 & 1.22 & 1.31 \end{pmatrix} \quad \text{and} \quad V_2 := \begin{pmatrix} 0.50 & 0.59 & 0.77 & 1.04 \\ 0.68 & 0.73 & 1.13 & 1.40 \\ 0.95 & 1.09 & 1.31 & 1.40 \end{pmatrix}$$

for E_1 and E_2, respectively. In mean, the variance is still 1 for both matrices V_1 and V_2. Under the null hypothesis, the methods are evaluated for the effect size matrix

$$E_0 := \begin{pmatrix} 0.0 & 0.1 & 0.3 & 0.6 \\ 0.2 & 0.2 & 0.3 & 0.6 \\ 0.5 & 0.5 & 0.5 & 0.6 \end{pmatrix},$$

which represents the same setting of the nuisance parameter as before and the effect $\vartheta_{ij} = \mu_{ij} - \max\{\mu_{i0}, \mu_{0j}\} = 0$ for all $(i, j) \in \mathbb{G}$. The type I error is evaluated for E_0 with equal variances over the treatment groups.
The results of the simulations are given in Table 4.6 for both the AVE- and MAX-test in the same parametric settings. For the normally distributed case, the bootstrap-based AVE-test performs slightly better than Hung's method if all effects are equal with equal variances. The power of the AVE-test is slightly greater for a design where the effects are distinct in the respective groups when applying the bootstrap to the lognormal distributed case. Under the null hypothesis, the significance level α is exceeded by Algorithm 4.3 for log-normal data. For the MAX-test, it turns out that the bootstrap approach has no greater power than the analytical method for any parametric and distributional setting. The simulations for the type I error of the MAX-test are equivalent to corresponding experiments on multiple min-tests under the complete null hypothesis: they reflect the probability for at least one true null hypothesis to be rejected by any min-test on the grid. The results for the MAX-test from Table 4.6 indicate that the familywise error is controlled weakly by the level $\alpha = 0.05$ as defined in (2.4).

4.3 Simultaneous confidence intervals

Confidence intervals are of common interest in clinical applications, where researchers are interested in possible values of outcome variables, supplementary to a rather technical adjusted p-value approach. The latter might particularly be useful for screening purposes. Simultaneous confidence intervals for k-factorial trials must satisfy the condition $P\left(\bigcap_{\gamma\in\mathbb{G}}\bigcap_{j=1}^{k}\left\{\mu_\gamma - \mu_{\gamma^j} \in I_{\gamma^j}\right\}\right) = 1-\alpha$. For construction of the intervals, the nuisance parameters δ_γ do not need to be considered because for a fixed treatment group $\gamma \in \mathbb{G}$, each of the k intervals describes the difference between the population mean μ_γ and only exactly one of the means $\mu_{\gamma^1}, ..., \mu_{\gamma^k}$. The calculation of confidence intervals therefore parallels the well-known problem of multiple intervals for a metric outcome and adequate contrast matrix.

			n=50			n=250
Population	Effect	Variance	B./G./H. (2001)	Algorithm 4.4	Algorithm 4.5	Algorithm 4.5
Normal	E_0	V_{hom}	0.9502	0.9634	0.9314	0.9442
	E_1	V_{hom}	0.9469	0.9658	0.9421	0.9442
		V_1	0.9191	0.9701	0.9288	0.9450
	E_2	V_{hom}	0.9461	0.9654	0.9308	0.9442
		V_2	0.9169	0.9629	0.9307	0.9435
Lognormal	E_0	V_{hom}	0.9422	0.9876	0.9140	0.9349
	E_1	V_{hom}	0.9440	0.9858	0.9140	0.9349
		V_1	0.9198	0.9452	0.8873	0.9249
	E_2	V_{hom}	0.9447	0.9835	0.9140	0.9338
		V_2	0.9104	0.9457	0.8868	0.9243

Table 4.6: *Simulation results for the coverage probabilities of the simultaneous confidence intervals are given for the various designs with calculations based on Algorithm 4.4, Algorithm 4.5 and the multivariate t-distribution as proposed by Bretz, Genz and Hothorn (2001).*

The most common approach is to determine the intervals by the representation

$$I_{\gamma^j} = \left[\bar{\mathbf{X}}_\gamma - \bar{\mathbf{X}}_{\gamma^j} - \xi\sqrt{\frac{\hat{\sigma}^2_\gamma}{n_\gamma} + \frac{\hat{\sigma}^2_{\gamma^j}}{n_{\gamma^j}}}, \bar{\mathbf{X}}_\gamma - \bar{\mathbf{X}}_{\gamma^j} + \xi\sqrt{\frac{\hat{\sigma}^2_\gamma}{n_\gamma} + \frac{\hat{\sigma}^2_{\gamma^j}}{n_{\gamma^j}}}\right]$$

with the same critical point ξ being used for all. Some of the manifold approaches to calculation of critical points in such a setting have been discussed in Chapter 2. Bretz, Genz and Hothorn (2001) proposed algorithms based on transformations of the multivariate t-distribution that allow numerical calculation of critical points with arbitrary precision for normally distributed and homoscedastic data. These methods are implemented in the `multcomp` package which is available on the Comprehensive R Archive Network (CRAN). Simultaneous confidence intervals for factorial designs are now constructed using the algorithm from Edwards and Berry (1987) which was introduced in Section 2.5.1. As mentioned in Chapter 2, the complete null hypothesis (3.15) is reflected in the resampling procedure. The critical points required for confidence interval estimation are obtained as the empirical $(1-\alpha)$-quantile of the test statistic $\max_{\gamma\in\mathbb{G}} |T^j_\gamma|$. Resampling is based on the residuals $\mathbf{X}_\gamma - \bar{\mathbf{X}}_\gamma$ using the respective sample variance estimators.

Algorithm 4.4 *(Simultaneous confidence intervals using test statistics)*

(1)	*Generate bootstrap samples* $\tilde{\mathbf{X}}^*_\gamma$ *from the residuals* $\mathbf{X}_\gamma - \bar{\mathbf{X}}_\gamma$ *for all* $\gamma \in \mathbb{G}_0$.	*Generate bootstrap samples* $\tilde{\mathbf{X}}^*_{ij}$ *from the residuals* $\mathbf{X}_{ij} - \bar{\mathbf{X}}_{ij}$ *for all* $(i,j) \in \mathbb{G}_0$.
(2)	*Calculate the statistics* T^{j*}_γ *for all* $\gamma \in \mathbb{G}$ *and* $j = 1, ..., k$ *and store* $\max_{\gamma \in \mathbb{G}, j=1,...,k} \lvert T^{j*}_\gamma \rvert$.	*Calculate the statistics* T^{A*}_{ij} *and* T^{B*}_{ij} *for all* $(i,j) \in \mathbb{G}$ *and store* $\max_{(i,j)\in\mathbb{G}} \{\lvert T^{A*}_{ij} \rvert, \lvert T^{B*}_{ij} \rvert\}$.
(3)	*Repeat steps (1) and (2)* N *times. Estimate the critical value* ξ *by the* m*-th order statistic of the* N *values of* $\max_{\gamma \in \mathbb{G}, j=1,...,k} \lvert T^{j*}_\gamma \rvert$ *with* $m = [(N+1)(1-\alpha)]$.	*Repeat steps (1) and (2)* N *times. Estimate the critical value* ξ *by the* m*-th order statistic of the* N *values of* $\max_{(i,j)\in\mathbb{G}} \{\lvert T^{A*}_{ij} \rvert, \lvert T^{B*}_{ij} \rvert\}$ *with* $m = [(N+1)(1-\alpha)]$.

These intervals are symmetric and still based on the test statistics which needs normal distributed data in both groups. The modification of Efron and Tibshirani's (1993) percentile method introduced in Section 2.5.1 can be used to overcome these limitations. For confidence intervals in k-factorial designs, this can be applied as denoted in the following algorithm.

Algorithm 4.5 *(Simultaneous confidence intervals using percentile method)*

(1)	*Generate bootstrap samples* $\tilde{\mathbf{X}}^*_\gamma$ *from* $\mathbf{X}_\gamma$ *for all* $\gamma \in \mathbb{G}_0$.	*Generate bootstrap samples* $\tilde{\mathbf{X}}^*_{ij}$ *from* $\mathbf{X}_{ij}$ *for all* $(i,j) \in \mathbb{G}_0$.
(2)	*Calculate the estimates* $\bar{\mathbf{X}}^*_\gamma - \bar{\mathbf{X}}^*_{\gamma^j}$ *for all* $\gamma \in \mathbb{G}$ *and* $j = 1, ..., k$. *Append these values as a row of a matrix* M. *Repeat (1) and (2)* N *times.*	*Calculate the estimates* $\bar{\mathbf{X}}^*_{ij} - \bar{\mathbf{X}}^*_{i0}$ *and* $\bar{\mathbf{X}}^*_{ij} - \bar{\mathbf{X}}^*_{0j}$ *for all* $(i,j) \in \mathbb{G}$ *and append these values as a row of a matrix* M. *Repeat (1) and (2)* N *times.*
(3)	*For* $l = 1, ..., (kD_1...D_k)$*, eliminate the rows of* M *with the smallest and largest value in the* l*th column. Repeat this step* $\frac{\alpha N}{2k}$ *times.*	*For* $l = 1, ..., 2AB$*, eliminate the rows of* M *with the smallest and largest value in the* l*th column. Repeat this step* $\frac{\alpha N}{2k}$ *times.*
(4)	*For the* l*th column, the estimates of the confidence limits for the corresponding comparison are* $\min_{1\le i\le k} M_{il}$ *and* $\max_{1\le i\le k} M_{il}$.	*For the* l*th column, the estimates of the confidence limits for the corresponding comparison are* $\min_{1\le i\le k} M_{il}$ *and* $\max_{1\le i\le k} M_{il}$.

The accuracy of the simultaneous intervals has been evaluated for the same data settings as the global hypothesis tests in the previous section. The simulation results in the last column of Table 4.5 represent simultaneous coverage probabilities of *all* intervals and are therefore desired to equal the nominal level of $1 - \alpha = 0.95$ closely. The simulations are given for the bootstrap approaches as well as for straight-forward calculation on multiple contrasts using the numerical methods for evaluation of multivariate t-integrals. For the latter, the coverage is close to the nominal level $1 - \alpha = 0.95$ for the homoscedastic case, but much lower for unequal variances in the involved treatment groups. Algorithm 4.4 performs slightly conservative for lognormal and homoscedastic data, but the coverage probability is at least $1 - \alpha$ in all cases. It can be shown by further simulations with different sample sizes that for $n < 50$ and even for $n = 5$, the nominal level is still protected but the intervals tend to be conservative if the samples are very small. For reasonable coverage probability, there should be a sample of at least $n = 15$. For Algorithm 4.5, the coverage is smaller than the nominal level in all cases if $n = 50$ but, however, approaches 0.95 for larger sample sizes as shown for $n = 250$ in the last column.

4.4 Combination drug for reduction of SiDBP: application to a clinical trial

For experiments on finding efficacious dose combinations in a k-factorial design, there are $(D_1 + 1) \cdot \ldots \cdot (D_k + 1)$ treatment groups to be tested and allocation of resources must be considered carefully. In particular, practical feasibility is limited to trials on disease patterns where clinical and financial expense for data acquisition are not to heavy. An example of such an experiment was given in the introduction. For convenience, it is reproduced here: a combination of a diuretic (drug A) and an ACE inhibitor (drug B) was tested for its efficacy in decrease of sitting diastolic blood pressure (SiDBP) with the response means and sample size allocation (in parentheses) summarized as follows:

	(A,0)	(A,1)	(A,2)	(A,3)
(B,0)	0 (75)	1.4 (75)	2.7 (74)	4.6 (48)
(B,1)	1.8 (74)	2.8 (75)	5.7 (74)	8.2 (49)
(B,2)	2.8 (48)	4.5 (50)	7.2 (48)	10.9 (48)

A pooled standard deviation of $\hat{\sigma} = 7.07$ was estimated. The results of Hellmich and Lehmacher (2005) for this example were summarized in the introduction. Some considerations on bootstrap-based simultaneous confidence intervals and adjusted p-values are now given to complement this. As the original data from the trial are kept confidential by the United States Food and Drug Administration (FDA), the calculations will be based

	normal homosc.			normal heterosc.			lognormal homosc.			lognormal heterosc.		
Contrast	p-value	2.5%	97.5%	p-value	2.5%	97.5%	p-value	2.5%	97.5%	p-value	2.5%	97.5%
11-10	0.6937	-2.288	4.288	0.6937	-2.286	4.286	0.6937	-2.287	4.287	0.6937	-2.289	4.289
11-01		-1.876	4.677		-1.875	4.675		-1.876	4.676		-1.879	4.678
12-10	0.0291	0.602	7.199	0.0291	0.603	7.197	0.0291	0.602	7.198	0.0291	0.600	7.200
12-02		-0.299	6.299		-0.297	6.297		-0.298	6.298		-0.300	6.300
13-10	0.0357	2.705	10.095	0.0357	2.706	10.094	0.0357	2.705	10.095	0.0357	2.703	10.097
13-03		-0.475	7.675		-0.473	7.673		-0.474	7.674		-0.477	7.677
21-20	0.5002	-2.354	5.754	0.5002	-2.352	5.752	0.5002	-2.354	5.754	0.5002	-2.357	5.757
21-01		-0.563	6.763		-0.561	6.761		-0.563	6.763		-0.565	6.765
22-20	0.0070	0.305	8.496	0.0070	0.307	8.494	0.0070	0.305	8.495	0.0070	0.302	8.498
22-02		0.782	8.218		0.783	8.217		0.782	8.218		0.780	8.220
23-20	4.8E-5	4.005	12.196	4.8E-5	4.007	12.194	4.8E-5	4.005	12.195	4.8E-5	4.002	12.198
23-03		2.205	10.396		2.207	10.394		2.205	10.395		2.202	10.398

Table 4.7: *Multiple inferences for simulated data matching the hypertension example from Hung (2000). For the p-values, the implementation of the unbalanced-design adjusted p-values approach was used (Hellmich and Lehmacher, 2005) where the calculations are based on the multivariate normal distribution (Chapter 3). The intervals are based on critical values of the multivariate t-distribution.*

on simulated samples all with the same descriptives as given in the above table, but with various distributional properties, i.e. the data are simulated from the normal and the lognormal distributions to analyze the behaviour for symmetric as well as for strongly skewed cases. For both types, one data set has been simulated with equal variances and one with linearly increasing standard deviations, i.e. the coefficient of variation for these data is held constant over the treatment groups.

The results of the calculations are summarized in Table 4.7 for the analytical approach based on the proposals of Hung (2000), Hellmich and Lehmacher (2005) and Bretz, Genz and Hothorn (2001) and in Table 4.8 for the bootstrap-based approach (Algorithms 4.2 and 4.4). Using the analytical methods, the intervals are essentially the same for all types of data because they are based on the corresponding sufficient statistics under additional distributional assumptions, that is to say homoscedasticity and normality of the data. In contrast, the alternative methods using Algorithm 4.2 for the adjusted p-values and Algorithm 4.4 for the simultaneous confidence intervals are sensitive to the actual distributional structure contained in the data. The width of the bootstrap-based intervals is different between the treatment groups not because of the unbalanced sample size allocation only but also due to the heterogenity of variances and different distributional shape of the data in use.

Consider the relation of the confidence intervals and the adjusted p-values as a kind of

	normal homosc.			normal heterosc.			lognormal homosc.			lognormal heterosc.		
Contrast	p-value	2.5%	97.5%	p-value	2.5%	97.5%	p-value	2.5%	97.5%	p-value	2.5%	97.5%
11-10	0.7280	-2.509	4.455	0.6455	-2.107	4.011	0.7315	-2.523	4.398	0.6228	-2.472	3.736
11-01		-2.907	4.844		-1.650	4.356		-2.111	4.787		-2.009	4.087
12-10	0.0327	0.379	7.367	0.0317	0.391	7.300	0.0278	0.366	7.310	0.0143	-0.022	6.991
12-02		-0.521	6.467		-0.606	6.495		-0.534	6.410		-1.030	6.176
13-10	0.0398	2.456	10.284	0.0890	1.893	10.767	0.0346	2.440	10.220	0.0576	1.363	10.370
13-03		-0.749	7.883		-1.529	8.571		-0.766	7.812		-2.133	8.118
21-20	0.5334	-2.628	5.961	0.4941	-2.544	5.812	0.5331	-2.644	5.891	0.4619	-3.043	5.438
21-01		-0.810	6.950		-0.648	6.732		-0.825	6.887		-1.089	6.401
22-20	0.0096	0.028	8.705	0.0134	-0.349	9.002	0.0072	-0.012	8.633	0.0046	-0.907	8.583
22-02		0.531	8.408		0.070	8.793		0.516	8.344		-0.452	8.402
23-20	2.0E-4	3.728	12.405	2.3E-3	2.684	13.348	1.0E-4	3.712	12.333	5.0E-4	2.047	12.870
23-03		1.928	10.605		0.668	11.758		1.912	10.533		0.006	11.261

Table 4.8: *Multiple inferences for simulated data matching the SiDBP example from Hung (2000). The calculations are based on Algorithm 4.2 for the p-values and Algorithm 4.4 for the intervals with $N = 25,000$ bootstrap iterations for each. The implementation in the* `bifactorial` *package was used that is available from the CRAN network.*

consistency control for the results. If both of the lower confidence bounds for $\mu_{ij} - \mu_{i0}$ and $\mu_{ij} - \mu_{0j}$ are non-negative, say

$$\min\left\{\bar{\mathbf{X}}_{ij} - \bar{\mathbf{X}}_{i0} - \xi\sqrt{\frac{\hat{\sigma}_{ij}^2}{n_{ij}} + \frac{\hat{\sigma}_{i0}^2}{n_{i0}}},\ \bar{\mathbf{X}}_{ij} - \bar{\mathbf{X}}_{0j} - \xi\sqrt{\frac{\hat{\sigma}_{ij}^2}{n_{ij}} + \frac{\hat{\sigma}_{0j}^2}{n_{0j}}}\right\} = 0,$$

this can likewise be denoted as $t_{ij}^{min} = \min\left\{t_{ij}^a,\ t_{ij}^b\right\} = \xi$, where the latter is equivalent to $\tilde{p}_{ij} = P\left(T_{ij}^{min} \geq t_{ij}^{min}\right) = \frac{\alpha}{2}$. As $1 - \alpha = 0.95$ is a common choice for the confidence level, the results are consistent if $\tilde{p}_{ij} \leq 0.025$ occurs if and only if both lower confidence bounds are non-negative.

Checking this in the results for the SiDBP example shows that the test decisions and intervals are consistent for the results based on the multivariate normal or t-distribution, whereas there are some abberations for heteroscedastic and lognormal data when using the bootstrap-based approach.

5 Binary endpoints in k-factorial designs

The outcomes of multiple dose experiments are often binary features, such as for example response vs. non-response to a medication, occurance vs. non-occurance of adverse events or recidive vs. non-recidive after a medicamentous therapy. Typically, the number of "events" in n cases is then binomially distributed with certain parameters π and n. Paralleling the test problem in (3.1), the null hypothesis of interest for the bifactorial design can then be expressed in terms of the proportion of "events" for patients in the respective treatment groups, that is to say

$$H_0^{ij} : (\pi_{ij} \leq \pi_{i0}) \vee (\pi_{ij} \leq \pi_{0j}) \quad \text{vs.} \quad H^{ij} : (\pi_{ij} > \pi_{i0}) \wedge (\pi_{ij} > \pi_{0j}), \tag{5.1}$$

representing the question if the event rate in the combination group is higher than in both component groups. To make inference on (5.1), the min-statistics can be based on appropriate χ^2- or Z-type statistics for comparisons of two proportions. In the following, the asymptotically normally distributed Z-statistics are chosen to point out the relationship to the continuous data case. The min-statistic is defined by $Z_{ij}^{min} = \min\left\{Z_{ij}^A, Z_{ij}^B\right\}$ with the denotations

$$Z_{ij}^A = \frac{\hat{\pi}_{ab} - \hat{\pi}_{a0}}{\sqrt{\frac{V(\hat{\pi}_{ab})}{n_{ab}} + \frac{V(\hat{\pi}_{a0})}{n_{a0}}}} \quad \text{and} \quad Z_{ij}^B = \frac{\hat{\pi}_{ab} - \hat{\pi}_{0b}}{\sqrt{\frac{V(\hat{\pi}_{ab})}{n_{ab}} + \frac{V(\hat{\pi}_{0b})}{n_{0b}}}}, \tag{5.2}$$

where $V(x) = x(1-x)$ and the $\hat{\pi}$ denote the sample "event" rates as estimates of the respective population rates. Wang and Hung (1997) proposed an approach to derive the power function of Z_{ij}^{min} for the special case of equal and large sample sizes in the treatment groups and discuss an extension to unequally-sized treatment groups. Again, the power function depends on the primary parameters $\vartheta_{ij} = \pi_{ij} - \max\{\pi_{i0}, \pi_{0j}\}$ and the (usually unknown) nuisance parameters $\delta_{ij} = \pi_{i0} - \pi_{0j}$. For monotonicity reasons, an upper bound for the type I error can be obtained by taking the supremum of the power function, evaluated at $\vartheta_{ij} = 0$ for all $(i,j) \in \mathbb{G}$, over all possible values of the nuisance parameters δ_{ij}. This offers a method to derive the p-values corresponding to any observed value of the test statistic, analogously as for the continuous data case.

Wang and Hung (1997) do not consider binary data designs with multiple dose combinations. This poses the problem of multiple inferences on hypothesis (5.1) using the min-test and requires considerations on the global question if *any* dose combination is more

efficacious than its components. The latter could be answered by global test methods for binary k-factorial designs comparable to those for the continuous data case. Derivations of the power functions of the corresponding AVE- and MAX-test and a multiple inference approach are not available up to now; approximations might involve the multivariate normal distribution for the Z statistics and a supremum taken over all possible values of the nuisance parameters δ_{ij}, which are now bounded between -1 and 1.
In the following section, the decisions on multiple dose combinations with binary endpoints will be based on resampling, generalizing the results from Chapter 4. Again, the methods are given for a grid where the dimensionality is not specified. The denotations $\mathbb{G} = \{1, ..., D_1\} \times ... \times \{1, ..., D_k\} \subset \mathbb{N}^k$, $\gamma = (i_1, ..., i_k)$ and $\gamma^j = (i_1, ..., i_{j-1}, 0, i_{j+1}, ..., i_k)$ are taken over from Chapter 3 and the null hypothesis is generalized as

$$H_0^\gamma : \bigvee_{j=1}^{k} (\pi_\gamma \leq \pi_{\gamma^j}) \quad \text{vs.} \quad H^\gamma : \bigwedge_{j=1}^{k} (\pi_\gamma > \pi_{\gamma^j}), \tag{5.3}$$

where the decision is based on the min-statistic $Z_\gamma^{min} = \min\{Z_\gamma^1, ..., Z_\gamma^k\}$ with the denotations

$$Z_\gamma^j = \frac{\hat{\pi}_\gamma - \hat{\pi}_{\gamma^j}}{\sqrt{\frac{V(\hat{\pi}_\gamma)^2}{n_\gamma} + \frac{V(\hat{\pi}_{\gamma^j})^2}{n_{\gamma^j}}}}, \tag{5.4}$$

analogously as for the t-statistics in (3.13). The Z_γ^j are approximately standard normal if the sample sizes are sufficiently high and the data are not too sparse.

5.1 Bootstrap approach

When the resampling-based approach is carried over from the continuous case (Chapter 4), some special problems arise for binary data. Regarding the first guideline of Hall and Wilson (1991), the null hypothesis should not be reflected by centering the data to a common mean like for continuous data. The result of this is not a $\{0, 1\}$-valued vector and its interpretation in connection with the "event" of interest is therefore unclear. Instead, either two samples of interest can be pooled to one single sample from which, subsequently, the random samples are drawn. The probabilities of an "event" are then

equal for both of those samples. Now, to reflect the null hypothesis

$$H_0^\gamma : (\pi_\gamma - \max\{\pi_{\gamma^1}, ..., \pi_{\gamma^k}\} = 0) \wedge \begin{pmatrix} \pi_{\gamma^1} - \pi_{\gamma^2} \\ \vdots \\ \pi_{\gamma^1} - \pi_{\gamma^k} \\ \pi_{\gamma^2} - \pi_{\gamma^3} \\ \vdots \\ \pi_{\gamma^2} - \pi_{\gamma^k} \\ \vdots \\ \pi_{\gamma^{k-1}} - \pi_{\gamma^k} \end{pmatrix} = \delta_\gamma, \tag{5.5}$$

how can the condition on the nuisance parameter vector δ_γ be involved when resampling from pooled data sets? The data are pooled to attain equal sample rates $\hat{\pi}_\gamma$ and $\max\{\hat{\pi}_{\gamma^1}, ..., \hat{\pi}_{\gamma^k}\}$. To reflect (5.5), the original distances of the event rates in the marginal groups should be kept when resampling the data from the remaining treatment groups, which can obviously not be attained by the pooling strategy.
As the *order* of the "events" in each treatment group is not relevant for the test and for resampling, it does not mean any loss of information on the data to involve the parameters $\hat{\pi}$ and n only, i.e. data are sampled from binomial distributions with the parameters $\big(\hat{\pi}_{\gamma^j}, n_{\gamma^j}\big)$ for $j = 1, ..., k$ and $\big(\max\{\hat{\pi}_{\gamma^1}, ..., \hat{\pi}_{\gamma^k}\}, n_\gamma\big)$. This reflects (5.5) and additionally offers a remarkable simplification in implementation.
The following is a particular form of what Efron and Tibshirani (1993) called a *parametric* bootstrap procedure. As before, the algorithms for binary data are given for general k-factorial designs as well as for $k = 2$.

Algorithm 5.1 *(Binary multiple min-tests with estimated nuisance parameters)*

(1)	*Initialize counting variables* $z_\gamma = 0$ *for* $\gamma \in \mathbb{G}$*. Calculate the min-statistics* Z_γ^{min} *following* (5.4).	*Initialize counting variables* $z_{ij} = 0$ *for* $(i, j) \in \mathbb{G}$*. Calculate the min-statistics* Z_{ij}^{min} *following* (5.2).
(2)	*Generate* $(D_1 + 1) \cdot ... \cdot (D_k + 1)$ *binomial samples* $\mathbf{X}_\gamma^*$ *that reflect the null hypothesis* (5.5)*: simulate* $\mathbf{X}_{\gamma^j}^*$ *from* $\mathbf{Bi}\big(\hat{\pi}_{\gamma^j}, n_{\gamma^j}\big)$ *for* $j = 1, ..., k$ *and* $\mathbf{X}_\gamma^*$ *from* $\mathbf{Bi}\big(\max\{\hat{\pi}_{\gamma^1}, ..., \hat{\pi}_{\gamma^k}\}, n_\gamma\big)$.	*Generate* $(A + 1) \cdot (B + 1)$ *binomial samples* $\mathbf{X}_{ij}^*$ *that reflect* (5.5)*: simulate* $\mathbf{X}_{i0}^*$ *from* $\mathbf{Bi}(\hat{\pi}_{i0}, n_{i0})$*,* $\mathbf{X}_{0j}^*$ *from* $\mathbf{Bi}(\hat{\pi}_{0j}, n_{0j})$ *and* $\mathbf{X}_{ij}^*$ *from* $\mathbf{Bi}(\max\{\hat{\pi}_{i0}, \hat{\pi}_{0j}\}, n_{ij})$.

	n=50		n=100	
(π_{A0}, π_{0B})	Algorithm 5.1	Algorithm 5.2	Algorithm 5.1	Algorithm 5.2
(0.9,0.7)	0.05560	0.05204	0.05080	0.04968
(0.8,0.6)	0.06244	0.04892	0.05424	0.05040
(0.7,0.5)	0.06428	0.04884	0.05576	0.05104
(0.6,0.4)	0.06872	0.05560	0.06124	0.05172
(0.5,0.3)	0.06376	0.04728	0.05456	0.04732
(0.4,0.2)	0.06220	0.04744	0.05172	0.04960
(0.3,0.1)	0.05780	0.04820	0.05288	0.05164

Table 5.1: *Results of* $25,000$ *simulations of the* 0.05*-level min-test for a binary outcome variable and* $\delta = 0.2$ *in a 1x1 bifactorial design using Algorithms 5.1 and 5.2 and* $N = 15,000$ *bootstrap iterations each. For both algorithms, the actual type I error slightly depends on the position of the rates of the marginal groups in the interval* $[0, 1]$*. This effect is stronger for* $n = 50$ *than for* $n = 100$*.*

(3)	*Calculate the min-test statistics* Z_γ^{min*} *from the resampled vectors* $\mathbf{X}_\gamma^*$ *and* $\{\mathbf{X}_{\gamma^j}^*\}_{j=1,\ldots,k}$ *following equation* (5.4)*. Check whether* $\max_{\gamma' \in \mathbb{G}} Z_{\gamma'}^{min*} \geq Z_\gamma^{min}$ *and in case increase* z_γ *by 1.*	*Calculate the min-test statistics* Z_{ij}^{min*} *from the resampled vectors* $\mathbf{X}_{ij}^*$*,* $\mathbf{X}_{i0}^*$ *and* $\mathbf{X}_{0j}^*$ *following equation* (5.2)*. Check whether* $\max_{(i',j') \in \mathbb{G}} Z_{i'j'}^{min*} \geq Z_{ij}^{min}$ *and in case increase* z_{ij} *by 1.*
(4)	*Repeat steps (2) and (3)* N *times and estimate the adjusted* p*-values by* $\hat{p}_\gamma^{(N)} = \frac{z_\gamma}{N}$*.*	*Repeat steps (2) and (3)* N *times and estimate the adjusted* p*-values by* $\hat{p}_{ij}^{(N)} = \frac{z_{ij}}{N}$*.*

Algorithm 5.1 has been evaluated for $k = 2$ in a design with only one combination group using a single min-test. When planning the simulation studies, several settings with $n = 50$ or $n = 100$ and a constant value of the nuisance parameter, that is to say $\delta = 0.2$, but distinct positions of the marginal rates in the interval $[0, 1]$ were simulated under the null hypothesis (Table 5.1). It turns out that the actual type I error is remarkably larger than 0.05 if the pair (π_{a0}, π_{0b}) is located in the mid-range of $[0, 1]$ but smaller near the limits of the interval. This a particular problem for binary data as in the continuous case, the null distribution of the test statistics does not depend on a location parameter. However, all simulation results in Table 5.1 are, even though not equal, but *close* to the nominal level $\alpha = 0.05$ especially for $n = 100$.

For the type I error, $25,000$ data sets were simulated for various parameter values δ and sample sizes n. For sake of clearness to the reader and because of the weak dependence on the position in $[0, 1]$, the simulations are performed for a fixed value of

Algorithm	n	δ=0.0	δ=0.1	δ=0.2	δ=0.3	δ=0.4
5.1	10	0.03264	0.05120	0.06768	0.07260	0.06548
	25	0.03440	0.06112	0.07348	0.06300	0.05176
	50	0.03144	0.06540	0.06428	0.05420	0.05164
	100	0.03272	0.07140	0.05576	0.05160	0.05208
	150	0.03360	0.07004	0.05268	0.04968	0.04996
	250	0.03308	0.06148	0.05060	0.05036	0.05036
5.2	10	0.07100	0.05732	0.06384	0.06860	0.07424
	25	0.01096	0.02552	0.04096	0.04636	0.04648
	50	0.01260	0.03612	0.04992	0.05128	0.05144
	100	0.01212	0.04336	0.05104	0.05192	0.05052
	150	0.01288	0.04716	0.04920	0.05268	0.04932
	250	0.01336	0.04884	0.05008	0.05008	0.05008

Table 5.2: *Results of $25,000$ simulations of the 0.05 per cent level min-test for a binary outcome variable in a 1x1 bifactorial design using Algorithms 5.1 and 5.2 and $N = 15,000$ bootstrap iterations each. For Algorithm 5.1, the type I error exceeds the nominal level if the nuisance parameter δ is slightly larger than zero. Using Algorithm 5.2, the test is very conservative for designs where the nuisance parameter δ is in an environment of zero. Note that a minimum amount of data should be available as for $n = 10$, the type I error exceeds the nominal level also when using Algorithm 5.2.*

$\pi_{a0} = 0.7$ and with various values $\pi_{0b} \in \{0.7, 0.6, 0.5, 0.4, 0.3\}$, which is equivalent to $\delta \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$. The simulation results are summarized in Table 5.2 and vizualised in Figure 5.1. Just as in the continuous data case, Algorithm 5.1 performs *anticonservative* for settings where δ is slightly larger than 0. In addition, the type I error depends on the sample size in a non-monotonic way which can be explained by the fact that the estimate of δ is more accurate for larger samples.

Again, a stricter assumption on the nuisance parameter is therefore proposed to achieve protection the significance level. Algorithm 5.2 is a modification of Algorithm 4.2 for the binary data case and implicitly assumes a large value for $|\delta|$.

Algorithm 5.2 *(Binary multiple min-tests with conservative assumption)*

(1)	*Initialize counting variables* $z_\gamma = 0$ *for* $\gamma \in \mathbb{G}$*. Determine the index* j *with* $\hat{\pi}_{\gamma^j} = \max\{\hat{\pi}_{\gamma^1}, ..., \hat{\pi}_{\gamma^k}\}$ *and define the min-statistics as* $Z_\gamma^{min} := Z_\gamma^j$.	*Initialize counting variables* $z_{ij} = 0$ *for* $(i,j) \in \mathbb{G}$ *and calculate the min-statistics* Z_{ij}^{min}.
(2)	*Generate* $(D_1+1) \cdot ... \cdot (D_k+1)$ *samples from the binomial distribution: simulate* $\mathbf{X}_{\gamma^j}^*$ *from* $\mathbf{Bi}(\hat{\pi}_{\gamma^j}, n_{\gamma^j})$ *for* $j = 1, ..., k$ *and* $\mathbf{X}_\gamma^*$ *from* $\mathbf{Bi}\big(\max\{\hat{\pi}_{\gamma^1}, ..., \hat{\pi}_{\gamma^k}\}, n_\gamma\big)$.	*Generate* $(A+1) \cdot (B+1)$ *samples from the binomial distribution: if* $\hat{\delta}_{ij} \geq 0$*, simulate* $\mathbf{X}_{ij}^*$ *from* $\mathbf{Bi}(\hat{\pi}_{i0}, n_{ij})$ *and* $\mathbf{X}_{i0}^*$ *from* $\mathbf{Bi}(\hat{\pi}_{i0}, n_{i0})$*, otherwise* $\mathbf{X}_{ij}^*$ *from* $\mathbf{Bi}(\hat{\pi}_{0j}, n_{ij})$ *and* $\mathbf{X}_{0j}^*$ *from* $\mathbf{Bi}(\hat{\pi}_{0j}, n_{0j})$.
(3)	*Determine the min-statistics* $Z_\gamma^{min*} := Z_\gamma^{j*}$. *Check whether* $\max_{\gamma' \in \mathbb{G}} Z_{\gamma'}^{min*} \geq Z_\gamma^{min}$ *and in case increase* z_γ *by 1.*	*Determine the min-statistics* Z_{ij}^{min*}. *Check whether* $\max_{(i',j') \in \mathbb{G}} Z_{i'j'}^{min*} \geq Z_{ij}^{min}$ *and in case increase* z_{ij} *by 1.*
(4)	*Repeat steps (2) and (3)* N *times. The adjusted* p*-value for group* γ *is then estimated by* $\hat{\tilde{p}}_\gamma^{(N)} = \frac{z_\gamma}{N}$.	*Repeat steps (2) and (3)* N *times. The adjusted* p*-value for group* (i,j) *is then estimated by* $\hat{\tilde{p}}_{ij}^{(N)} = \frac{z_{ij}}{N}$.

This has been evaluated in the same way as Algorithm 5.1. Again, the type I error slightly depends on the position of the marginal means in the interval $[0, 1]$; thus in Table 5.1, the simulation results for various positions in $[0, 1]$ are given also for Algorithm 5.2. The simulation results are shown in the second panel of Table 5.2: the algorithm performs conservative for small values of the nuisance parameter δ, but the actual type I error is approximately $\alpha = 0.05$ if δ is greater than 0, i.e. the behaviour is comparable to that of Algorithm 4.2. Note that if the sample size is extremely small ($n = 10$), the significance level is exceeded also with this conservative method.

Algorithm 5.2 has been evaluated also under the alternative hypothesis: data were generated from the binomial distribution and the population variance was determined as $n\pi(1-\pi)$ from the parameters π and n. Thus, assumptions on distributional shape and equal or unequal variances over the respective groups do not make sense. As the range for combinations of the parameters δ and ϑ is limited by the restriction that all parameters have to be in the interval $[0, 1]$, the simulations are performed for a fixed value $\pi_{a0} = 0.5$ with $\pi_{0b} \in \{0.5, 0.4, 0.3, 0.2, 0.1\}$, which is equivalent to $\delta \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$. In addition, the values $\pi_{ab} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ are considered, representing a primary parameter of $\vartheta \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$. The results of the power simulations are given in Table 5.3, showing a similar behaviour as for the continuous case as reported in Table

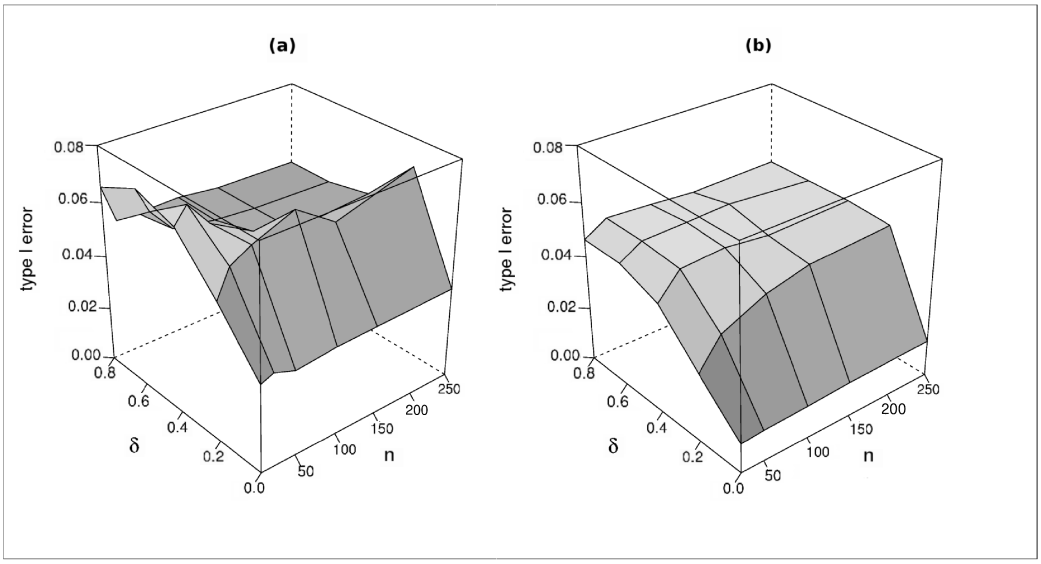

Figure 5.1: *Vizualisation of the results in Table 5.2.* ***(a)*** *Simulations are based on Algorithm 5.1 where the nominal level $\alpha = 0.05$ is exceeded for nuisance parameters $\delta > 0$.* ***(b)*** *Evaluation of Algorithm 5.2: the type I error rate is much smaller than $\alpha = 0.05$, particularly if δ is in an environment of 0. The significance level is exceeded also for Algorithm 5.2 if the sample size is very small ($n = 10$), which has been omitted in the figure to get a clearer vizualisation.*

4.3: the power is monotonically increasing in both parameters δ and ϑ. The procedure tends to a more conservative behaviour if δ is an environment around zero, but the power is satisfying also in these cases if, in addition, the effect size is sufficiently large.

5.2 Global null hypotheses for binary data

The global question if *any* combination group evocates a higher response than both of its components is now extended to the binary case. Following the remarks in Section 3.3, the null hypothesis and alternative of interest can be expressed according to

$$H_0 : \forall \gamma \in \mathbb{G} : \bigvee_{j=1}^{k} (\pi_\gamma \leq \pi_{\gamma^j}) \quad \text{vs.} \quad H_1 : \exists \gamma \in \mathbb{G} : \bigwedge_{j=1}^{k} (\pi_\gamma > \pi_{\gamma^j}). \tag{5.6}$$

ϑ	$\delta = 0.00$	$\delta = 0.10$	$\delta = 0.20$	$\delta = 0.3$
0.00	0.0144	0.0337	0.0429	0.0515
0.05	0.0504	0.1009	0.1189	0.1243
0.10	0.1376	0.2301	0.2599	0.2635
0.15	0.3089	0.4319	0.4557	0.4589
0.20	0.5348	0.6511	0.6804	0.6839
0.25	0.7588	0.8389	0.8496	0.8576
0.30	0.9013	0.9433	0.9448	0.9493

Table 5.3: *Simulation results for the power of the* 0.05 *level min-test for a 1x1 bifactorial design using Algorithm 5.2. The power is monotonically increasing in* δ *also for non-vanishing values of the primary parameter* ϑ.

For the decision on this test problem, the AVE- and MAX-statistics are carried over from the continuous case. Paralleling the AVE- and MAX-tests proposed by Hung, Chi and Lipicky (1993) and Hung (2000), the generalized form

$$Z_{ave} = (D_1 ... D_k)^{-1} \sum_{\gamma \in \mathbb{G}} Z_\gamma^{min} \quad \text{and} \quad Z_{max} = \max_{\gamma \in \mathbb{G}} Z_\gamma^{min}$$

will be used. A modification of the AVE-test algorithm from Section 4.2 can be applied to the case of binary data. The null hypothesis will be reflected in the same way as mentioned for the multiple inference case in Section 5.1: data are sampled from the binomial distribution with parameters $\left(\hat{\pi}_{\gamma^j}, n_{\gamma^j}\right)$ for $j = 1, ..., k$ and $\left(\max\{\hat{\pi}_{\gamma^1}, ..., \hat{\pi}_{\gamma^k}\}, n_\gamma\right)$, respectively. From these, the AVE- and MAX-statistics are calculated. The algorithm for the AVE-test is the following.

Algorithm 5.3 *(Binary resampling-based AVE-test)*

(1)	*Initialize a counting variable* $z = 0$. *Determine the min-statistics* Z_γ^{min} *and calculate the average statistic* $Z_{ave} = (D_1 \cdot ... \cdot D_k)^{-1} \sum_{\gamma \in \mathbb{G}} Z_\gamma^{min}$ *from the data.*	*Initialize a counting variable* $z = 0$. *Determine the min-statistics* Z_{ij}^{min} *and calculate the average statistic* $Z_{ave} = (AB)^{-1} \sum_{(i,j) \in \mathbb{G}} Z_{ij}^{min}$ *from the data.*
(2)	*Generate* $(D_1 + 1) \cdot ... \cdot (D_k + 1)$ *samples from the binomial distrubution: simulate* $\mathbf{X}_{\gamma^j}^*$ *from* $\mathbf{Bi}\left(\hat{\pi}_{\gamma^j}, n_{\gamma^j}\right)$ *for* $j = 1, ..., k$ *and* $\mathbf{X}_\gamma^*$ *from* $\mathbf{Bi}\left(\max\{\hat{\pi}_{\gamma^1}, ..., \hat{\pi}_{\gamma^k}\}, n_\gamma\right)$.	*Generate* $(A + 1) \cdot (B + 1)$ *samples from the binomial distribution: if* $\hat{\delta}_{ij} \geq 0$, *simulate* $\mathbf{X}_{ij}^*$ *from* $\mathbf{Bi}\left(\hat{\pi}_{i0}, n_{ij}\right)$ *and* $\mathbf{X}_{i0}^*$ *from* $\mathbf{Bi}\left(\hat{\pi}_{i0}, n_{i0}\right)$, *otherwise* $\mathbf{X}_{ij}^*$ *from* $\mathbf{Bi}\left(\hat{\pi}_{0j}, n_{ij}\right)$ *and* $\mathbf{X}_{0j}^*$ *from* $\mathbf{Bi}\left(\hat{\pi}_{0j}, n_{0j}\right)$.

Effect size	AVE-test	MAX-test	Confidence intervals
E_0	0.0607	0.0272	0.9548
E_3	0.7390	0.5752	0.9545
E_4	0.9375	0.6148	0.9796

Table 5.4: *Simulation results for the* 0.05 *level AVE- and MAX-tests based on* $E3$ *and* $E4$ *taken as the effect size matrices. The calculations are based on Algorithms 5.3 and 5.2, where* $25,000$ *simulations were performed under the null hypothesis and* $10,000$ *for* E_3 *and* E_4*. The overall coverage probability of simultaneous confidence intervals based on Algorithm 5.4 has also been evaluated for either setting of treatment effects. For all analyses,* $N = 15,000$ *bootstrap iterations were used.*

(3)	*Determine the min-test statistics* $Z_\gamma^{min*} = Z_\gamma^{j*}$ *and the average statistic* $Z_{ave}^* = (D_1 \cdot \ldots \cdot D_k)^{-1} \sum_{\gamma \in \mathbb{G}} Z_\gamma^{min*}$*. Increase* z *by one if* $Z_{ave}^* \geq Z_{ave}$*.*	*Determine the min-test statistics* Z_{ij}^{min*} *and the average statistic* $Z_{ave}^* = (AB)^{-1} \sum_{(i,j) \in \mathbb{G}} Z_{ij}^{min*}$*. Increase* z *by one if* $Z_{ave}^* \geq Z_{ave}$*.*
(4)	*Repeat steps (2) and (3)* N *times. The estimated* p*-value is then* $\hat{p}_{ave}^{(N)} = \frac{z}{N}$*.*	*Repeat steps (2) and (3)* N *times. The estimated* p*-value is then* $\hat{p}_{ave}^{(N)} = \frac{z}{N}$*.*

The p-value for the MAX test can be derived from the multiple inference method as given by Algorithm 5.2, where p_{max} is estimated by the smallest of the resulting p-values.

The performance of Algorithm 5.3 and of the MAX-test based on Algorithm 5.2 have been evaluated by simulation experiments. The matrix

$$E_0 = \begin{pmatrix} 0.0 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.2 & 0.3 & 0.4 \\ 0.3 & 0.3 & 0.3 & 0.4 \end{pmatrix}$$

represents the complete null hypothesis, i.e. the effect sizes for all $(i,j) \in \mathbb{G}$ have the common value $\vartheta_{ij} = \pi_{ij} - \max\{\pi_{i0}, \pi_{0j}\} = 0$, whereas

$$E_3 = \begin{pmatrix} 0.0 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.4 & 0.5 & 0.6 \\ 0.3 & 0.5 & 0.5 & 0.6 \end{pmatrix} \quad \text{and} \quad E_4 = \begin{pmatrix} 0.00 & 0.20 & 0.30 & 0.40 \\ 0.20 & 0.25 & 0.55 & 0.65 \\ 0.30 & 0.55 & 0.55 & 0.65 \end{pmatrix}$$

now indicate the respective response rates in the combination treatment groups when simulating under the alternative. Paralleling the simulation in Chapter 4, the matrix

E_3 represents a design with an effect of 0.2 in all combination groups, whereas E_4 has smaller effects in the low-dose group $(1,1)$ but the mean of the rates in all groups is still 0.2. The variances in the respective groups are determined from these effects through the representation $n\pi(1-\pi)$ as the data are generated from the binomial distribution. A balanced sample size allocation of $n = 100$ per group was chosen.
The results are given in Table 4.5. It turns out that the behaviour is similar to that of the methods for the continuous data application: the resampling-based AVE-test tends to exceed the significance level under the null hypothesis, whereas the MAX-test performs more conservative than the AVE-test in all situations. As the latter is equivalent to an evaluation of the multiple inference procedure, it is important to remark that the familywise error is protected weakly by the significance level α.

5.3 Confidence intervals for binary data

For binary data, confidence intervals for the differences in the "event" rates between the combination groups and their respective components, i.e. for $\pi_\gamma - \pi_{\gamma^j}$ and $j = 1, ..., k$, can be given. As pointed out in Section 4.3, no problems arise from the nuisance parameters δ_γ because the group γ and only one of the groups γ^j, $j = 1, ..., k$, is involved in each inference. Covering all parameters simultaneously with a prespecified probability $1-\alpha$ means $P\left(\bigcap_{\gamma\in\mathbb{G}}\bigcap_{j=1}^{k}\left\{\pi_\gamma - \pi_{\gamma^j} \in I_\gamma^j\right\}\right) = 1-\alpha$ to hold for the intervals, where

$$I_\gamma^j = \left[\hat{\pi}_\gamma - \hat{\pi}_{\gamma^j} - \xi\sqrt{\frac{V(\hat{\pi}_\gamma)}{n_\gamma} + \frac{V(\hat{\pi}_{\gamma^j})}{n_{\gamma^j}}}, \hat{\pi}_\gamma - \hat{\pi}_{\gamma^j} + \xi\sqrt{\frac{V(\hat{\pi}_\gamma)}{n_\gamma} + \frac{V(\hat{\pi}_{\gamma^j})}{n_{\gamma^j}}}\right]. \tag{5.7}$$

Hence the remaining issue is to estimate the critical value ξ. As it is no loss of information to consider the parameters π and n for each sample only, data are sampled from the binomial distributions $\mathbf{Bi}\left(\frac{\hat{\pi}_\gamma+\hat{\pi}_{\gamma^j}}{2}, n_\gamma\right)$ and $\mathbf{Bi}\left(\frac{\hat{\pi}_\gamma+\hat{\pi}_{\gamma^j}}{2}, n_{\gamma^j}\right)$, reflecting the null hypothesis $\pi_\gamma = \pi_{\gamma^j}$ for $j = 1, ..., k$ and $\gamma \in \mathbb{G}$. Summarizing these considerations, the algorithm can be outlined as follows.

Algorithm 5.4 *(Binary simultaneous confidence intervals)*

(1)	*Generate* $2kD_1 \cdot ... \cdot D_k$ *binary samples* $\{(\mathbf{X}^*_\gamma, \mathbf{X}^*_{\gamma^j})\}_{\gamma\in\mathbb{G};j=1,...,k}$ *from the binomial distributions* $\mathbf{Bi}\Big(\frac{\hat\pi_\gamma+\hat\pi_{\gamma^j}}{2}, n_\gamma\Big)$ *and* $\mathbf{Bi}\Big(\frac{\hat\pi_\gamma+\hat\pi_{\gamma^j}}{2}, n_{\gamma^j}\Big)$.	*Generate* $4AB$ *binary samples* $\{(\mathbf{X}^*_{ij}, \mathbf{X}^*_{i0}), (\mathbf{X}^*_{ij}, \mathbf{X}^*_{0j})\}_{(i,j)\in\mathbb{G}}$ *from the binomial distributions* $\mathbf{Bi}\Big(\frac{\hat\pi_{ij}+\hat\pi_{i0}}{2}, n_{ij}\Big)$, $\mathbf{Bi}\Big(\frac{\hat\pi_{ij}+\hat\pi_{i0}}{2}, n_{i0}\Big)$, $\mathbf{Bi}\Big(\frac{\hat\pi_{ij}+\hat\pi_{0j}}{2}, n_{ij}\Big)$, *and* $\mathbf{Bi}\Big(\frac{\hat\pi_{ij}+\hat\pi_{0j}}{2}, n_{0j}\Big)$.
(2)	*Calculate the statistics* Z^{j*}_γ *for all* $\gamma \in \mathbb{G}$ *and* $j = 1, ..., k$ *as given by* (5.4). *Store the value* $\max_{\gamma\in\mathbb{G},j=1,...,k} \lvert Z^{j*}_\gamma\rvert$.	*Calculate the statistics* Z^{A*}_{ij} *and* Z^{B*}_{ij} *for all* $(i, j) \in \mathbb{G}$ *as given in* (5.2). *Store the value* $\max_{(i,j)\in\mathbb{G}}\{\lvert Z^{A*}_{ij}\rvert, \lvert Z^{B*}_{ij}\rvert\}$.
(3)	*Repeat steps (1) and (2)* N *times. Estimate the critical value* ξ *by the* m*-th order statistic of the* N *values of* $\max_{\gamma\in\mathbb{G},j=1,...,k} \lvert Z^{j*}_\gamma\rvert$ *with* $m = [(N + 1)(1 - \alpha)]$.	*Repeat steps (1) and (2)* N *times. Estimate the critical value* ξ *by the* m*-th order statistic of the* N *values of* $\max_{(i,j)\in\mathbb{G}}\{\lvert Z^{A*}_{ij}\rvert, \lvert Z^{B*}_{ij}\rvert\}$ *with* $m = [(N + 1)(1 - \alpha)]$.

The simultaneous coverage probability of the intervals constructed by Algorithm 5.4 has been evaluated by an additional simulation experiment for the same designs as for the global AVE- and MAX-test. The results for the confidence intervals are given in the last column of Table 5.4 together with those for the AVE- and MAX-tests: under the null hypothesis and in the setting E_3, the coverage probability is close to the nominal level 0.95. For the design with unequal effect sizes, the intervals still cover the true parameters with sufficient probability but are slightly more conservative.

5.4 Remission of AML patients under combined decitabine and cytarabine therapy

As a binary data application, the trial supervised by Huang et al. (2007) reported in the introduction is now analyzed. For convenience, the example is reproduced here: patients suffering from acute myeloid leukemia (AML) are treated by a combination of the two drugs decitabine and cytarabine. The binary feature to be a responder or non-responder is taken as the endpoint of this trial where the response criterion is taken to be achievement of complete remission. The response rates are expected to look approximately like

	(A,0)	(A,1)	(A,2)
(B,0)	–	0.45 (31)	0.65 (17)
(B,1)	0.30 (100)	0.71 (50)	0.70 (50)
(B,2)	0.59 (101)	0.64 (50)	0.75 (50)

Contrast	p-value	2.5%	97.5%
11-10	0.0417	0.1714	0.6456
11-01		-0.0819	0.5843
12-10	0.8208	0.1480	0.6287
12-02		-0.3686	0.4355
21-20	0.7434	-0.2196	0.2869
21-01		-0.1679	0.5118
22-20	0.5848	-0.0807	0.3898
22-02		-0.3020	0.4895

Table 5.5: *Multiple inferences for simulated data matching the specifications in the AML remission example. The p-values have been determined by Algorithm 5.1, whereas Algorithm 5.4 was used for the confidence intervals. The implementation is available in the* `bifactorial` *package.*

with the respective sample sizes given in parentheses. Application of the bootstrap approach to this requires a complete binary data set for which the descriptive statistics coincide with the values in the above table. These can be simulated randomly on a personal computer. Multiplicity-adjusted p-values according to the min-test, confidence intervals and p-values for the AVE- and MAX-test were determined for the example and summarized in Table 5.5. Combination $(1,1)$ has got the desired property that the response rate is significantly higher than in both component groups. Note that the confidence intervals are consistent with the min-test p-values as discussed in section (4.4): for the $(1,1)-(0,1)$ contrast, the lower confidence bound is close but still smaller than zero, whereas the p-value is slightly larger than 0.025.
The AVE-test based on Algorithm 5.3 results in $p_{ave} = 0.0523$. As discussed in Chapter 3, the MAX-test p-value is determined as the smallest of the results from the multiple hypothesis approach, i.e. $p_{max} = 0.0416$. These p-values both indicate that at least one combination is better than both of its components, although the AVE-test does not show significance. In the multiple procedure, the desirable combination was identified to be the drug in group $(1,1)$.

6 Discussion

Several bootstrap-based approaches to the min-test and corresponding global test statistics on k-factorial clinical trial designs have been proposed and evaluated in terms of statistical power. However, as mentioned in Section 2.5.4, the level of uncertainty represented by the confidence limits of the simulation results should be kept in mind throughout the discussion. First, the method based on Algorithm 4.1 was observed to exceed the nominal level α if the nuisance parameters are slightly greater than zero. This problem analogously occurs in the binary data case (Algorithm 5.1) and emerges from the fact that except for reasonably large sample sizes, the estimate $\hat{\delta}$ will most likely not match the true value δ and resampling under the null hypothesis is therefore possible in an approximate sense only. This is now considered on a probabilistic level: the min-statistic can be written as $T_{min} = T_A \mathbb{1}_{\{T_A \leq T_B\}} + T_B \mathbb{1}_{\{T_B < T_A\}}$ and has got the expectation value

$$\mathrm{E}[T_{min}] = \mathrm{E}\left[T_A \mathbb{1}_{\{T_A \leq T_B\}}\right] + \mathrm{E}\left[T_B \mathbb{1}_{\{T_A > T_B\}}\right].$$

For $\delta \gg 0$, the random variable $\mathbb{1}_{\{T_A \leq T_B\}}$ reduces to the constant 1 as $P\left(\{T_A \leq T_B\}\right) \approx 1$. An analogous conclusion holds for $\delta \ll 0$ such that

$$\mathrm{E}[T_{min}] = \begin{cases} \mathrm{E}[T_A] & \text{if} \quad \delta \gg 0 \\ \mathrm{E}[T_B] & \text{if} \quad \delta \ll 0. \end{cases} \tag{6.1}$$

For settings where the approximation $|\delta| = \infty$ is not valid, the expectation value of T_{min} can be derived applying rules that are commonly known from probability theory, i.e.

$$\begin{aligned} \mathrm{E}\left[T_A \mathbb{1}_{\{T_A \leq T_B\}}\right] &= \int_{-\infty}^{\infty} \sum_{t=0}^{1} st \; \varphi^{(T_A, \mathbb{1}_{\{T_A \leq T_B\}})}(s,t) \; ds \\ &= \int_{-\infty}^{\infty} s \; \varphi^{(T_A, \mathbb{1}_{\{T_A \leq T_B\}})}(s,1) \; ds \end{aligned}$$

In the latter, the common p.d.f. of T_A and $\mathbb{1}_{\{T_A \leq T_B\}}$ is denoted by $\varphi^{(T_A, \mathbb{1}_{\{T_A \leq T_B\}})}$ and can be written as

$$\varphi^{(T_A, \mathbb{1}_{\{T_A \leq T_B\}})}(s,1) = P(T_B \geq T_A | T_A = s) P(T_A = s) = (1 - F^{T_B}(s)) f^{T_A}(s)$$

where F^{T_B} denotes the c.d.f. of the test statistic T_B and f^{T_A} the p.d.f. of T_A. For the above expectation value, it follows now that

$$\mathrm{E}\big[T_A \mathbb{1}_{\{T_A \leq T_B\}}\big] = \int_{-\infty}^{\infty} s\,(1 - F^{T_B}(s)) f^{T_A}(s)\,ds.$$

Analogously, the second summand of $\mathrm{E}[T_{min}]$ can be shown to have the representation $\mathrm{E}\big[T_B \mathbb{1}_{\{T_B \leq T_A\}}\big] = \int_{-\infty}^{\infty} s\,(1 - F^{T_A}(s)) f^{T_B}(s)\,ds$. Taken together, these results give the expectation value of the min-statistic as a function of the distributions of T_A and T_B:

$$\mathrm{E}[T_{min}] = \int_{-\infty}^{\infty} s\left((1 - F^{T_B}(s)) f^{T_A}(s) + (1 - F^{T_A}(s)) f^{T_B}(s)\right) ds. \tag{6.2}$$

Under the null hypothesis $H_0 : (\mu_{ab} - \max\{\mu_{a0}, \mu_{0b}\} = 0) \wedge (\mu_{a0} - \mu_{0b} = \delta)$, the test statistics T_A and T_B are t-distributed with $n_{ab} + n_{a0} - 2$ and $n_{ab} + n_{0b} - 2$ degrees of freedom and noncentrality parameters

$$\mu_A = \begin{cases} 0 & \text{for } \delta > 0 \\ -\delta / \sqrt{\frac{1}{n_{ab}} + \frac{1}{n_{a0}}} & \text{for } \delta \leq 0 \end{cases} \quad \text{and} \quad \mu_B = \begin{cases} \delta / \sqrt{\frac{1}{n_{ab}} + \frac{1}{n_{0b}}} & \text{for } \delta > 0 \\ 0 & \text{for } \delta \leq 0. \end{cases}$$

Thus, the expectation value $\mathrm{E}[T_{min}]$ can be obtained numerically by equation (6.2). If Algorithm 4.1 is used for resampling, the bootstrapped mean $\hat{\mu}_{min}$ of the distribution of T_{min} therefore depends on the estimated value $\hat{\delta}$ and does in general not equal the population mean μ_{min} of the min-statistic under the null hypothesis. Now, the shape of this dependence is not symmetric and as the probabilities of δ to be over- or underestimated are approximately equal, the resulting sample mean is more likely to be $\hat{\mu}_{min} < \mu_{min}$ than $\hat{\mu}_{min} \geq \mu_{min}$ if $|\delta| > 0$. In total, there are less iterations where $T^*_{min} \geq T_{min}$ than would be under the true null hypothesis, biasing the p-values towards smaller values, i.e. anticonservative test results.
Figure 6.1 shows the result of $5,000$ simulations of a 2x2-design with $\delta = 0.3$ and $n = 50$, where estimates of δ were explicitly evaluated and plotted against the bootstrap distributional mean $\hat{\mu}_{min}$ of the min-test T_{min}, resulting from $N = 15,000$ bootstrap iterations each. The mean $\hat{\mu}_{min}$ does highly depend on the estimated value $\hat{\delta}$ in an asymmetric way. It tends to be smaller for $\hat{\delta} < \delta$ than for $\hat{\delta} > \delta$. It can be shown by further simulations for choices of δ other than 0.3 that the shape of the dependence is nearly independent from the true value of δ.

Simulations of binary data applications using Algorithm 5.1 cause slightly different problems than for the continuous case and Algorithm 4.1. Despite the nuisance parameter

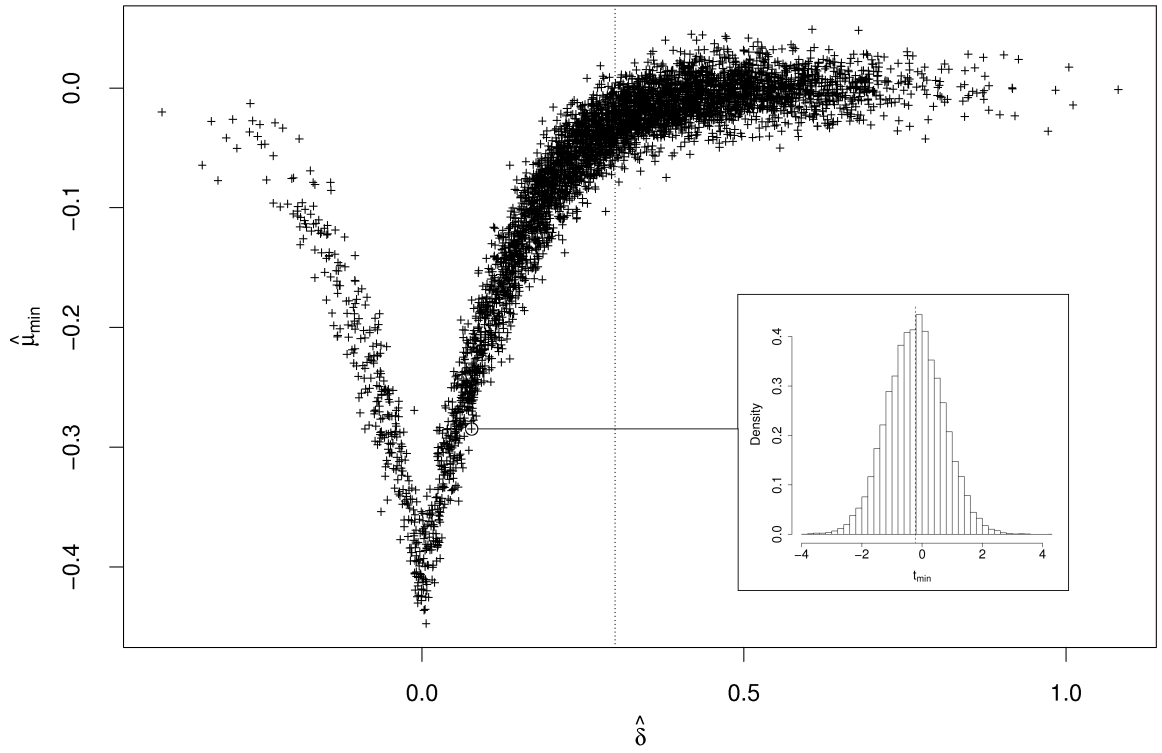

Figure 6.1: *Each of the* $5,000$ *dots represents a single sample from the null distribution with* $\delta = 0.3$ *and* $n = 50$ *in all three cells of a 2x2 design. The empirical distribution of the test statistic was calculated by* $N = 15,000$ *bootstrap iterations per sample and is exemplified on the right for one of the* $5,000$ *replicates. The plot shows the dependence of the resulting distributional mean* $\hat{\mu}_{min}$ *on the estimated nuisance parameter* $\hat{\delta}$*. Small values are more likely to occur when* $\hat{\delta} < \delta$ *than for* $\hat{\delta} > \delta$*. The dotted line represents the true value* $\delta = 0.3$*.*

$\delta = \pi_{a0} - \pi_{0b}$, the actual type I error slightly depends on the position of the parameters π_{a0} and π_{0b} in the interval $[0, 1]$. The deviations from the nominal level result from the poor estimation of δ in the small-sample case; note that this problem is almost not present for $n = 100$. It is no surprise that the performance of Algorithm 5.1 is similar to that of Algorithm 4.1 as the test statistics are approximately normally distributed also in the binary case, following from the Central Limit Theorem. On the other hand, there is only a finite number of possible simulated data sets and thus for the nuisance parameter: for $n = 50$, there are $51 \times 51 = 2601$ pairs of binary vectors, resulting in 101 different values of δ. For vizualisation, $2,500$ binary data sets have been simulated and the mean $\hat{\mu}_{min}$ of the sampled distribution of the min-statistic Z_{min} based on $N = 15,000$ bootstrap

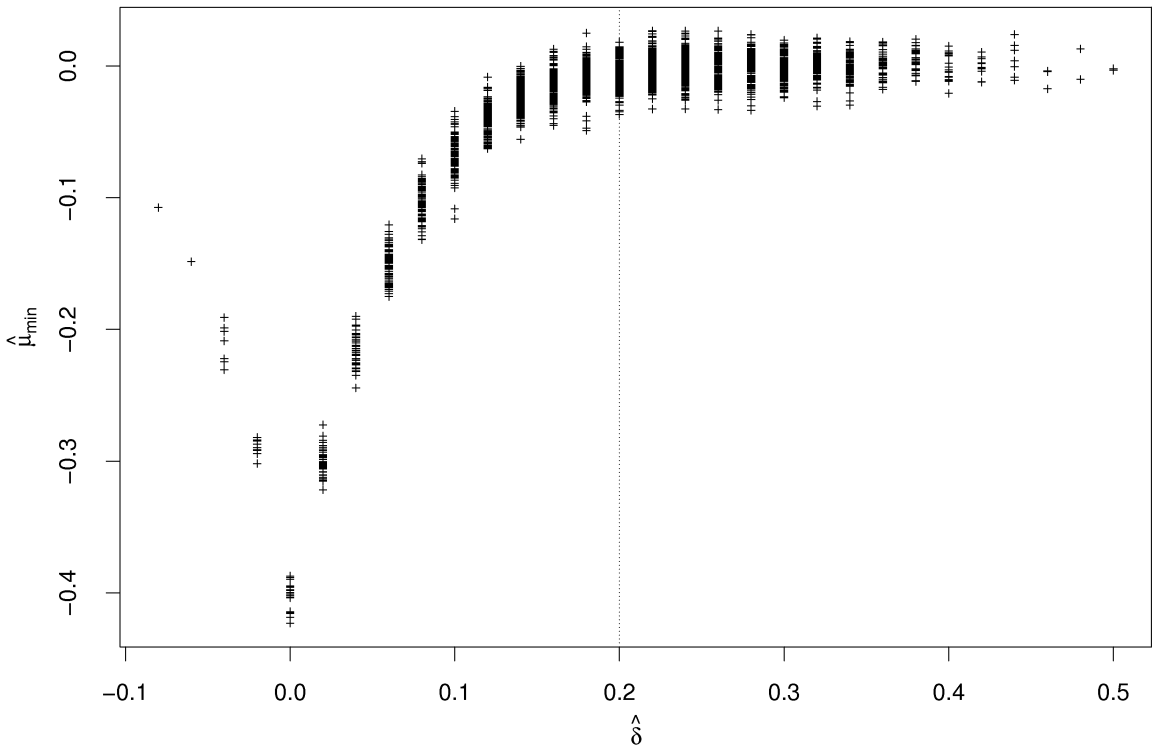

Figure 6.2: *Dependence of the sampled distributional mean $\hat{\mu}_{min}$ on the estimated nuisance parameter $\hat{\delta}$ in the binary case. Each of the $2,500$ dots represents a single sample from the null distribution with $\delta = 0.2$ and $n = 50$ in all three cells of a binary 2x2 design. The empirical distribution was calculated by $N = 15,000$ bootstrap iterations for each sample. The shape of the dependence is similar as in the continuous case.*

iterations has been plotted against the estimated nuisance parameter in each simulation (Figure 6.2). Note the difference between this and the continuous case plotted in Figure 6.1. Thus the accuracy of the estimate for the nuisance parameter is not sufficient also in the binary case. The considerations on the distributional mean of the min-statistic depending on the estimate $\hat{\delta}$ also apply to binary data as the expectation of the min-statistic has been derived without any assumptions on the type of the statistic that the min-test is based on; thus it can be directly converted to

$$\mathrm{E}\left[Z_{min}\right] = \int_{-\infty}^{\infty} (1 - F^{Z_B}(s)) f^{Z_A}(s) + (1 - F^{Z_A}(s)) f^{Z_B}(s) ds \tag{6.3}$$

where F^{Z_A}, F^{Z_B} denote the c.d.f. of the test statistics Z_A, Z_B and f^{Z_A}, f^{Z_B} the corresponding density functions.

The dependence of $\mathrm{E}[T_{min}]$ on the estimated value $\hat{\delta}$ is a general problem which also arises in further approaches involving estimated values for the nuisance parameters: Snapinn (1987) reports that anticonservative p-values occur when handling with several estimates for the nuisance parameters based on the data observed, with $|\hat{\delta}| = \infty$ being the only assumption that uniformly keeps the significance level. The conclusion from this is that without knowledge of δ, the min-test cannot essentially be improved, particularly for settings where δ is close to zero.

The simulations for $k = 3$ give information on the behaviour of the min-test for higher dimensionality. The type I error was supposed to depend primarily on the proportion of the largest mean values in the marginal groups. The $\binom{k}{2}$-dimensional nuisance parameter vector δ is uniquely determined by $k-1$ of his components. Denote by $\mu_{(1)} \leq ... \leq \mu_{(k)}$ the marginal group means $\mu_{\gamma^1}, ..., \mu_{\gamma^k}$ in increasing order. The components of δ can then be represented as the $k-1$ values

$$|\mu_{(k)} - \mu_{(k-1)}| \leq ... \leq |\mu_{(k)} - \mu_{(1)}| \tag{6.4}$$

Under the null hypothesis, these determine which is the smallest of the k statistics $T^1_\gamma, ..., T^k_\gamma$ involved in the min-test: if $|\mu_{(k)} - \mu_{(k-1)}|$ is large, it is very unlikely that any other marginal group than that with mean $\mu_{(k)}$ yields the smallest test statistic. If on the other hand, $|\mu_{(k)} - \mu_{(k-1)}|$ is comparatively small, the probability that the group with mean $\mu_{(k-1)}$ yields the smallest statistic is larger. In addition, if the next component $|\mu_{(k)} - \mu_{(k-2)}|$ is only slightly larger, it is more likely that the marginal group with mean $\mu_{(k-2)}$ yields the smallest statistic. Thus, the $k-1$ values in (6.4) have monotonically decreasing impact on the test level where, importantly, the dependence of the level is strongest for the smallest of these values. In addition, for any value in the list, the influence of those above it is smaller the larger that particular value is.

This interpretation is in accordance with the simulation results in the last four columns of Table 4.2. For $k = 3$, the simulation experiments showed that the distance $|\mu_{(3)} - \mu_{(2)}|$ is in fact the marginal parameter that has strongest influence on the observed type I error as the latter is distinct for the marginal means $(0.5, 0.4, 0.0)$ and $(0.5, 0.1, 0.0)$, whereas the nuisance parameter vector is the same but in different order. Considering the results for $(0.5, 0.4, 0.4)$, the probability that the test involving $\mu_{(3)}$ and $\mu_{(1)}$ is smallest is higher than in the $(0.5, 0.4, 0.0)$ case, whereas this probability is essentially equal for $(0.5, 0.1, 0.1)$ and $(0.5, 0.1, 0.0)$ as the test statistic for the first marginal group is always smaller. Summarizing these results, the actual type I error does not depend on the parameters $|\mu_{(k-1)} - \mu_{(k-2)}|, ..., |\mu_{(2)} - \mu_{(1)}|$ if $|\mu_{(k)} - \mu_{(k-1)}|$ is large enough. If, on the other hand, the latter is close to zero, the dependence on the next parameter $|\mu_{(k-1)} - \mu_{(k-2)}|$ becomes

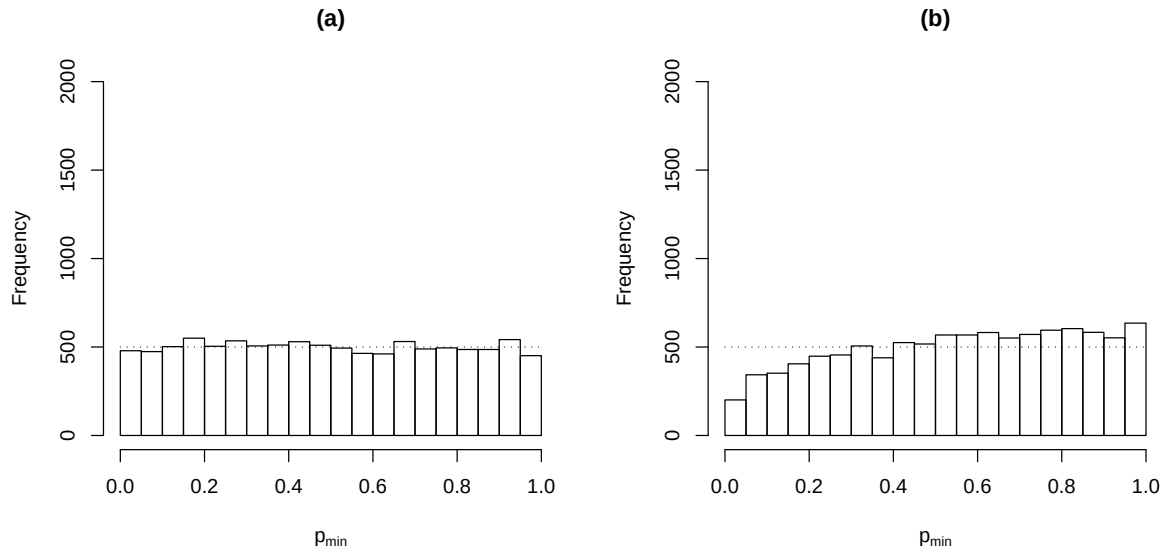

Figure 6.3: *Frequency of p-values for the min-test under the null hypothesis with nuisance parameter* ***(a)*** $\delta = 0.5$ *and* ***(b)*** $\delta = 0.1$ *using* $20,000$ *samples from a normal population with* $n = 50$ *in each marginal group. The proportion of values below* 0.05 *is smaller for the case with* $\delta = 0.1$*, whereas the histogram for* $\delta = 0.5$ *is the same as for a two-sample* t*-test.*

stronger. Now, the distribution of the min-statistic under the null hypothesis cannot be generated without knowledge of $|\mu_{(k)} - \mu_{(k-1)}|$ also for general choices of k; i.e. when using Algorithms 4.1 or 5.1, also the general null hypothesis (4.1) can be reflected in an approximate sense only unless the $|\mu_{(k)} - \mu_{(k-1)}|$ can be determined properly. For $k = 2$, the only way out of this was obtained to be the assumption $|\delta| = \infty$ which is used by Algorithm 4.2 and causes a more conservative performance. This conclusion reflects the fact that in the case of large nuisance parameters, the min-test is an ordinary two-sample t-test and the p-values of the single statistics are therefore rectangular-shaped for normally distributed data (Figure 6.3a). If, on the other hand, $|\delta|$ is in an environment of zero, the min-statistic is the minimum of two t-statistics the p-values for which are both uniformly distributed. For the min-test with the t-distribution taken as the reference distribution, larger p-values are therefore more likely to occur (Figure 6.3b). Both taken together, this is the reason why the nominal level is matched well for large $|\delta|$, whereas the min-test tends to be very conservative for $|\delta| \approx 0$ when using Algorithm 4.2. Concluding from the $k = 2$ case to generality, the type I error is expected to be close to α if the assumption $|\mu_{(k)} - \mu_{(k-1)}| = \infty$ is approximately satisfied which is confirmed by the

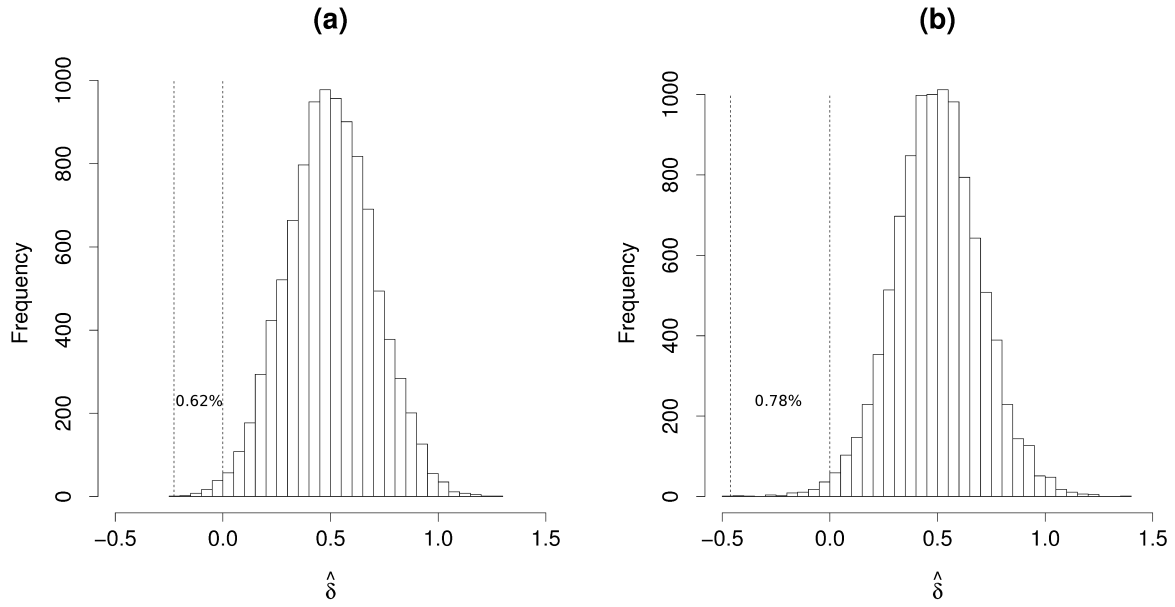

Figure 6.4: *Frequency of estimated values for a true nuisance parameter* $\delta = 0.5$ *using* $10,000$ *samples from* ***(a)*** *a normal and* ***(b)*** *a lognormal population with* $n = 50$ *in each marginal group. The proportion of values smaller than zero is larger for the lognormal case, thus values* $\hat{\delta}$ *with a sign that is distinct from that of* δ *are more frequent here. The sections on the axes where this occurs are marked by dashed lines in both figures.*

results for $k = 3$.
In the simulations for skewed populations and large values of $|\delta|$, it was found that the probability to reject a true null hypothesis is still larger than the nominal level α if 4.2 is used. This is due to the fact that in this algorithm, the *sign* of the nuisance parameter is still determined from the data. For skewed populations, the proportion of mismatches in the sign of δ and thus in determining the min-statistic is higher than for the symmetric case. Results from a simulation of normal and lognormal data are vizualized in Figure 6.4 and show the sections on the axes where the described errors occur. If $|\delta|$ is large, rejecting a true null hypothesis is more likely if the wrong comparison is chosen for the min-statistic. Thus the type I error is in fact larger than α in the lognormal case.

The confidence intervals based on the multivariate t-distribution have a coverage probability remarkably less than $1-\alpha = 0.95$ if heteroscedasticity of the populations is ignored. It is not possible to involve the sample variances from the data in that kind of analysis,

and as instead equal variances are assumed throughout the analysis, the resulting intervals all have the same length if the sample sizes in the groups are equal. Thus, the event that mean differences in certain groups are not covered happens too often such that the coverage probability is smaller than 0.95 in these cases.
In contrast, the actual sample distribution is used in a natural way when the bootstrap is applied. Regarding the results from the hypertension example in Tables 4.7 and 4.8, it becomes obvious how the improvement in terms of coverage works: the bounds of the simultaneous confidence intervals are adjusted according to the respective group variances only when using the bootstrap, but not for the multivariate t-distribution. Nevertheless, the coverage of the bootstrap intervals (Algorithm 4.4) may become conservative for certain parametric assumptions but is at least $1 - \alpha$ in all cases.
For the new approach to simultaneous confidence intervals based on the percentile method (Algorithm 4.4), it turned out that the coverage probability is not sufficient for moderate sample sizes ($n = 50$) but very close to the nominal level of 0.95 for any distributional setting if $n = 250$. Following from the Glivenko-Cantelli theorem, the bootstrap distribution of the parameter estimates forms a better approximation to the empirical c.d.f. if the sample size is large. For $n = 50$, the approximation is not sufficiently smooth particularly in the distributional tails, causing a high uncertainty about the confidence limits. However, as the performance is satisfying for larger sample sizes, the principle of Algorithm 4.5 is in working order and may be improved by a mechanism for the number of iterations to omit on both sides of the interval, depending on the sample size available.

Evaluation of the bootstrap-based AVE- and MAX-test showed that the respective power is not essentially better than if the analytical methods of Hung (2000) are applied. The behaviour of these methods can be explained by the fact that the bootstrap-based p-values do not coincide entirely with those determined by analytical methods also for the multiple min-tests as discovered, for instance, in the example of Section 4.4. These aberrations are reproduced and amplified in the global methods derived from the min-test. Nevertheless, for binary data applications, the bootstrap-based AVE- and MAX-tests form the only existing approach to the global null hypothesis in this special problem. Recommendations of analysis methods for the global hypotheses which are most suitable in terms of power were given by Buchheister (2001) for several formations of the effect sizes. In this context, the performance of the bootstrap-based AVE- and MAX-tests is comparable to that of Hung's tests with the above-mentioned limitations.

The discussion is concluded with some remarks concerning problems of factorial designs

that have not been considered here and form remaining lacks in the factorial design theory. For many reasons it may make sense to omit certain combination groups from the design, one of which is that these combinations are expected to achieve little effect. Patients undergoing these treatments would otherwise be consciously excluded from a more efficacious therapy. Hung (1996) showed that the power of the test procedures is then remarkably lower compared to a complete design. Buchheister (2001) proposed to solve these problems by the closed test principle, but a bootstrap approach could also be appropriate to allow for empty groups in the factorial design grid.

The interpretation of the results for the $k = 3$ case has been mostly qualitative in this scope, bringing up the question how the analytical basis of this approach looks, particularly how the power functions of the tests depend on the components of the nuisance parameter vector ordered by size as in (6.4). The result of this might also be extended to the unbalanced case as done by Hung (2000) for the $k = 2$ case. Furthermore, literature on analytical approaches to bifactorial designs involving binary data is very limited up to now.

Further research is needed for the extension of the theory to group sequential or adaptive designs. Ethical and logistic issues as well as financial problems that arise from the possibly very large total sample size needed for a factorial trial design are weakened when applying adaptive designs where the recruitment in certain groups can be stopped for futility at an early stage after interim analyses. Lehmacher, Kieser and Hothorn (2000) applied group sequential and adaptive methods to multiple testing problems. This may help saving an essential number of patients to be included in the trial, but also leads to the problem of incomplete factorial designs mentioned above.

Apart from efficacy analysis, drug safety and the evaluation of side effects are important issues in any clinical trial. These have not been considered here as application of the min-test and its related methods to this is not suggested by the legal situation. Safety-related events are typically reported descriptively but can also be analyzed by the factorial design methods, leading to the question if particular side effects in a combination drug are at most as frequent as in its respective components.

7 Summary

Bifactorial designs are used to test for the efficacy of drugs with a fixed combination of two (or more) components. The question if a combination has a significantly higher efficacy than both of its components is of common interest for obvious clinical reasons and due to legal prescriptions. For this, Laska and Meisner (1989) proposed the min-test which is defined by the minimum of two statistics, each testing the difference in efficacy between the combination and one of the single compounds. To give an answer to the question if there is *at least one* combination with the desired property, Hung, Chi und Lipicky (1993) proposed the AVE- and MAX-test where the average or the maximum, respectively, of the involved min-statistics is taken. This was initially stated for the balanced and homoscedastic case. Hung (2000) generalized the theory to designs with unequally-sized groups. The question *which* of the combinations have got the property mentioned above involves the problem of multiple hypotheses testing.
The power of these methods highly depends on the difference of the respective parameters between the component groups considered in the analysis. This is described by the so-called nuisance parameters δ_{ij} that quantify the marginal differences between the combination groups: if these parameters are close to zero, the methods are very conservative, whereas they are more powerful for large values.
In general, bootstrap methods are a suitable approach to simulate the distribution of test statistics without the need for an analytical representation. However, it turns out that the dependence on the nuisance parameters δ_{ij} is in the nature of the min-statistic and is still present if the reference distribution is approximated by a bootstrap algorithm. Furthermore, estimating the δ_{ij} from the data does not give any improvement as the simulated distribution of the min-statistic asymmetrically depends on the estimated value, the accuracy of which is quite uncertain, at least for small samples. Nevertheless, the bootstrap methods offer additional advantages: no assumptions on the distribution of the data as e.g. a normality assumption or homogenity of the variances in the respective groups are needed. Furthermore, it is not relevant if the sample sizes in the groups are equal, and the correlation of the test statistics is automatically included in the calculation because data which *two* particular statistics are based on are used for both throughout the simulation of the distribution. In contrast to an analytical approach, the extension

of the methods to designs for combinations of more than two compounds requires no further theory.

The power of the proposed bootstrap algorithms was evaluated by simulation experiments. Basically, only one single min-test was considered under the corresponding null hypothesis and the dependency on the nuisance parameter and sample size was analyzed. When estimating the marginal difference from the data, it was shown that for $|\delta| > 0$, the nominal level of $\alpha = 0.05$ is exceeded. If the conservative assumption $|\delta| = \infty$ is used instead as implicitly done in the approach of Hung (2000), the type I error is substantially smaller than the significance level for $\delta \approx 0$. In principle, the problem connected to the unknown nuisance parameters can therefore not be resolved by bootstrap methods. The power of the latter conservative method has additionally been evaluated for various values of n for purpose of sample size planning.

Using bootstrap algorithms, the AVE- and MAX-tests can be performed with an easier implementation than in the classical approach as analytical considerations on the distribution functions are needless for these. The behaviour in simulation experiments was similar to that of Hung's approach in most situations. Additionally, confidence intervals can be given with a comparatively simple implementation using bootstrap methods. In connection to bifactorial designs, nothing of this kind is available from literature up to now. The coverage probability was simulated using a bootstrap approach involving the test statistics as well as for a novel algorithm for bootstrap percentiles in multiple problems, considering various distributional conditions for both methods. It was shown that for the bootstrap approach involving the test statistics, the coverage probability is always at least $1 - \alpha = 0.95$ in contrast to a procedure based on the multivariate t-distribution. However, for skewed distributions, the bootstrap intervals are quite conservative. The bootstrap percentile intervals for multiple settings have sufficient coverage probability only for large sample sizes. Considering the hypertension trial from the paper of Hung (2000), classical methods result in the same intervals for unequal variances as for homoscedastic data because the pooled variance estimate is always used. On the other hand, the bootstrap methods yield different intervals the length of which depends on the respective sample variance.

All the methods described above have been transferred to the case of binary endpoints for which theory is hardly available from literature. However, the results essentially agree to those for the continuous case because of asymptotical properties of the test statistics (Central Limit Theorem). For the continuous case, the min-test for generalized dimensionality was evaluated for combinations of 3 compounds. For this, it was shown that the power primarily depends on the distance of the two marginal groups with the largest

population means and the other marginal means have an influence only if this parameter is very small. If the distances are estimated from the data, the nominal significance level is no more protected if the value of the parameter is in fact positive. Using the assumption that both means are far apart from each other, the test is very conservative also in the three-dimensional case. In the discussion, it was shown that this can be generalized to arbitrary dimensionality from which the tests for combinations of 2 or 3 component drugs are obtained as special cases.

8 Zusammenfassung

Bifaktorielle Studienpläne werden verwendet, um die Wirksamkeit von Präparaten mit einer festen Kombination aus zwei (oder mehr) Wirkstoffen zu testen. Aus naheliegenden klinischen Gründen und aufgrund regulatorischer Vorgaben interessiert man sich für die Frage, ob eine Kombination eine signifikant höhere Wirksamkeit aufweist als beide Komponenten für sich betrachtet. Laska und Meisner (1989) schlugen hierfür den min-Test vor, dessen Teststatistik durch das Minimum von zwei Statistiken gebildet wird, die die Wirksamkeit der Kombination gegen je eines der Einzelpräparate testen. Um die Frage zu beantworten, ob es *mindestens eine* Kombination mit der gewünschten Eigenschaft gibt, schlugen Hung, Chi und Lipicky (1993) den AVE- und MAX-Test vor, bei dem der Mittelwert bzw. das Maximum der beteiligten min-Statistiken gebildet wird. Dies ist zunächst nur für den balancierten und homoskedastischen Fall geschehen. Hung (2000) verallgemeinerte diese Theorie auf Designs mit ungleichen Gruppengrößen. Die Frage, *welche* der betrachteten Kombinationen das o.g. Kriterium erfüllen, führt in die Problematik des multiplen Testens.
Man stellt fest, dass die Mächtigkeit der beschriebenen Methoden in hohem Maß davon abhängt, wie stark sich die jeweiligen Parameter zwischen den betrachteten Gruppen der Einzelpräparate unterscheiden. Dies wird durch die sogenannten Störparameter δ_{ij} beschrieben, die die Randdifferenzen der Parameter zwischen den Gruppen der Einzelpräparate angeben: Liegen diese Parameter nahe Null, so sind die Methoden sehr konservativ, während sie für große Werte erheblich mächtiger sind.
Bootstrap-Methoden sind im Allgemeinen geeignet, um die Verteilung von Teststatistiken zu simulieren, ohne eine analytische Darstellung zu benötigen. Es zeigt sich aber, dass die o.g. Abhängigkeit von den Störparametern δ_{ij} in der Natur der min-Statistik liegt und daher auch noch besteht, wenn die Prüfverteilung mit einem Bootstrap-Algorithmus approximiert wird. Auch das Schätzen der δ_{ij} aus den Daten führt zu keiner Verbesserung, weil die simulierte Verteilung der min-Statistik in asymmetrischer Weise von dem geschätzten Wert abhängt, der zumindest für kleine Stichproben mit großer Unsicherheit behaftet ist. Dennoch haben die verwendeten Bootstrap-Methoden noch weitere Vorteile: Es fließen keine Annahmen über die Verteilung der Daten ein wie z. B. eine Normalverteilungsannahme oder die Homogenität der Varianzen in den Gruppen. Ferner spielt es keine Rolle,

ob die Stichproben in den Gruppen verschieden groß sind, und die Korrelation der Teststatistiken fließt automatisch in die Berechnung ein, weil von zwei Statistiken gemeinsam verwendete Daten auch in der Simulation der Verteilung gemeinsam verwendet werden. Es ist im Gegensatz zu einem analytischen Ansatz ohne weiteres möglich, die Verfahren auf Designs für Kombinationspräparate mit mehr als zwei Wirkstoffen auszudehnen.
Die vorgeschlagenen Bootstrap-Algorithmen wurden in Simulationsexperimenten auf ihre Mächtigkeit hin untersucht. Dabei wurde zunächst nur ein einzelner min-Test unter der entsprechenden Nullhypothese betrachtet und die Abhängigkeit vom Störparameter und der Stichprobengröße analysiert. Wenn die Randdifferenz aus den Daten geschätzt wurde, zeigte sich dabei, dass für $|\delta| > 0$ das nominale Niveau von hier $\alpha = 0.05$ deutlich überschritten wurde. Arbeitet man hingegen mit der konservativen Annahme $|\delta| = \infty$, so wie dies implizit im Ansatz von Hung (2000) geschieht, so wird für $\delta \approx 0$ das Niveau nicht einmal annähernd ausgeschöpft. Die Problematik der unbekannten Störparameter lässt sich also mit den untersuchten Bootstrap-Methoden prinzipiell nicht auflösen. Die erwähnte konservative Methode wurde zum Zweck der Fallzahlplanung auch für verschiede Werte von n ausgewertet.
Die AVE- und MAX-Tests können durch die vorschlagenen Methoden mit sehr viel übersichtlicherer Methodik als klassisch durchgeführt werden, weil eine analytische Betrachtung der Verteilungsfunktionen sich erübrigt. Das Verhalten in Simulationsexperimenten war in den meisten Situationen ähnlich wie mit der von Hung untersuchten Technik. Konfidenzintervalle können durch Bootstrap-Methoden ebenfalls mit einfacher Implementation angegeben werden. Dergleichen wurde im Zusammenhang mit faktoriellen Studienplänen bisher noch nicht in der Literatur beschrieben. Die Überdeckungswahrscheinlichkeit wurde sowohl mit einem auf den Teststatistiken basierenden Algorithmus als auch für eine neue Methode untersucht, bei der die multiplen Intervallgrenzen als Bootstrap-Perzentile bestimmt werden. Beide Varianten wurden für unterschiedliche Verteilungsbedingungen simuliert und es konnte gezeigt werden, dass die Wahrscheinlichkeit bei der erstgenannten Methode im Gegensatz zu einem auf der multivariaten t-Verteilung basierenden Verfahren immer mindestens $1 - \alpha = 0.95$ beträgt, allerdings bei schiefgipfligen Verteilungen recht konservativ wird. Die Überdeckungswahrscheinlichkeit der durch Bootstrap-Perzentile konstruierten Intervalle ist nur für größere Stichprobenumfänge ausreichend. Am Beispiel der Bluthochdruckstudie aus dem Artikel von Hung (2000) zeigte sich, dass klassische Methoden im Fall ungleicher Varianzen die gleichen Intervalle ergeben wie bei homoskedastischen Daten, weil dort immer der gepoolte Varianzschätzer verwendet wird. Die Bootstrap-Methoden erzeugen hingegen Intervalle, deren Breite von der jeweiligen Stichprobenvarianz abhängt.

Alle erwähnten Verfahren wurden auf den Fall binärer Endpunkte übertragen, zu dem bislang nur wenig in der Literatur zu finden ist (Wang und Hung, 1997). Allerdings sind die Ergebnisse hier aufgrund asymptotischer Eigenschaften der Teststatistiken (Zentraler Grenzwertsatz) weitgehend identisch mit denen für den stetigen Fall. Für diesen wurde der min-Test außerdem mit verallgemeinerter Dimensionalität für Präparate mit 3 Wirkstoffen ausgewertet. Hierbei zeigte sich, dass die Mächtigkeit des Verfahrens maßgeblich vom Abstand der beiden Randgruppen mit den jeweils größten Mittelwerten abhängt und die anderen Ränder nur eine Rolle spielen, wenn dieser Parameter sehr klein ist. Schätzt man diesen Abstand aus den Daten, so wird das nominale Signifikanzniveau wiederum deutlich überschritten, wenn der wirkliche Parameter einen positiven Wert hat. Verwendet man immer die Annahme, dass die beiden Mittelwerte weit voneinander entfernt sind, so ist der Test auch für den dreidimensionalen Fall sehr konservativ. In der Diskussion wurde gezeigt, dass sich dies auf beliebige Dimensionalität verallgemeinern lässt, worin die Tests für Präparate mit 2 oder 3 Wirkstoffen als Spezialfälle enthalten sind.

9 References

[1] Babu GJ, Singh K (1983): *Inference on means using the bootstrap.* The Annals of Statistics **11(3)**, 999-1003

[2] Bretz F, Genz A, Hothorn L (2001): *On the numerical availability of multiple comparison procedures.* Biometrical Journal **43**, 645-656

[3] Buchheister B (2001): *Statistische Methoden zum Nachweis der Effektivität von Kombinationspräparaten.* PhD thesis, University of Cologne

[4] Buchheister B, Lehmacher W (2006): *Multiple testing procedures for identifying desirable dose combinations in bifactorial designs.* GMS Medizinische Informatik, Biometrie und Epidemiologie **2(2)**, Doc07

[5] Committee for Proprietary Medicinal Products, Efficacy Working Party (2002): *Points to consider on multiplicity issues in clinical trials, CPMP/EWP/908/99.* Available at http://www.emea.eu.int/pdfs/human/ewp/090899en.pdf

[6] Edwards D, Berry JJ (1987): *The efficiency of simulation-based multiple comparisons.* Biometrics **43**, 913-928

[7] Efron B (1979): *Bootstrap methods: Another look at the jackknife.* The Annals of Statistics **7(1)**, 1-26

[8] Efron B, Tibshirani RJ (1993): *An introduction to the bootstrap.* Chapman & Hall, New York

[9] Gabriel KR (1970): *On the relation between union-intersection and likelihood ratio tests.* Essays in Probability and Statistics, University of North Carolina Press, 251-266

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

[12] Genz A, Bretz F (2002): *Comparison of methods for the computation of multivariate t-probabilities.* Journal of Computational and Graphical Statistics **11**, 950-971

[13] Hall P, Wilson SR (1991): *Two guidelines for bootstrap hypothesis testing.* Biometrics **47**, 757-762

[14] Hayter AJ (1984): *A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative.* The Annals of Statistics **12**, 61-75

[15] Hellmich M, Lehmacher W (2005): *Closure procedures for monotone bifactorial dose-response designs.* Biometrics **61**, 269-276

[16] Hochberg Y, Tamhane AC (1987): *Multiple comparison procedures.* John Wiley & Sons, Inc., New York

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

[18] Hsu JC (1996): *Multiple comparisons: Theory and methods.* Chapman & Hall, New York

[19] Huang X, Biswas S, Oki Y, Issa JP, Berry DA (2007): *A parallel phase I/II clinical trial design for combination therapies.* Biometrics **63(2)**, 429-436

[20] Hung HMJ, Chi GYH, Lipicky RJ (1993): *Testing for the existence of a desirable dose combination.* Biometrics **49**, 85-94

[21] Hung HMJ (1994): *Testing for the existence of a desirable dose combination (Correspondence).* Biometrics **50**, 307-308

[22] Hung HMJ (1996): *Global tests for combination drug studies in factorial trials.* Statistics in Medicine **15**, 233-248

[23] Hung HMJ (2000): *Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials.* Statistics in Medicine **19**, 2079-2087

[24] Issa JPJ, Garcia-Manero G, Giles FJ, Mannari R, Thomas D, Faderl S, Bayar E, Lyons J, Rosenfeld CS, Cortes J and Kantarjian HM (2004): *Phase I study of low-dose prolonged exposure schedules of the hypomethylating agent 5-aza-2'-deoxycytidine (decitabine) in hematopoietic malignancies.* Blood **103(5)**, 1635-1640.

[25] Jogdeo K (1977): *Association and probability inequalities.* The Annals of Statistics **5**, 495-504

[26] Kramer CY (1956): *Extension of multiple range tests to group means with unequal number of replications.* Biometrics **12**, 307-310

[27] Laska EM, Meisner MJ (1989): *Testing whether an identified treatment is best.* Biometrics **45**, 1139-1151

[28] Lehmacher W, Kieser M, Hothorn L (2000): *Sequential and multiple testing for dose-response analysis.* Drug Information Journal **34**, 591-597

[29] Marcus R, Peritz E, Gabriel KR (1976): *On closed testing procedures with special reference to ordered analysis of variance.* Biometrika **63**, 655-660

[30] Petersdorf SH, Rankin C, Head DR, Terebelo HR, Willman CL, Balcerzak SP, Karnad AB, Dakhil SR, Appelbaum FR (2007): *Phase II evaluation of an intensified induction therapy with standard daunomycin and cytarabine follwed by high-dose cytarabine for adults with previously untreated acute myeloid leukemia: A southwest oncology group study (SWOG-9500)*. American Journal of Hematology **82(12)**, 1056-62

[31] Roy SN (1953): *On a heuristic method of test construction and its use in multivariate analysis.* The Annals of Mathematical Statistics **24**, 220-238

[32] Shao J (1999): *Mathematical Statistics.* Springer, New York

[33] Šidák Z (1967): *Rectangular rejection regions for the means of multivariate normal distributions.* Journal of the American Statistical Association **62**, 626-633

[34] Singh K (1981): *On the asymptotic accuracy of Efron's bootstrap.* The Annals of Statistics **9(6)**, 1187-1195

[35] Snapinn SS (1987): *Evaluating the efficacy of a combination therapy.* Statistics in Medicine **6**, 657-665

[36] Tukey JW (1953): *The problem of multiple comparisons.* Unpublished manuscript

[37] Wang SJ, Hung HMJ (1997): *Large sample tests for binary outcomes in fixed-dose combination drug studies.* Biometrics **53**, 498-503

[38] Westfall PH, Young SS (1993): *Resampling-based multiple testing.* John Wiley & Sons, Inc., New York

10 Appendix: Implementation of the algorithms

Manifold approaches to factorial designs have been proposed by several authors. Some advantages and disadvantages of the various methods were discussed in the preceding chapters. To make all the methods available for practitioners, they have been summarized to an R package called `bifactorial` that is available from the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/web/packages/bifactorial/index.html`. The package offers construction of class `carpet` and `cube` objects that are appropriate to represent bi- or trifactorial designs. Graphics and descriptive statistics can be generated as well as the inductive methods such as the AVE- and MAX-test. For inference on multiple hypotheses, the algorithms for calculation of adjusted p-values and simultaneous confidence intervals for the treatment groups involved in the design have been implemented.
Method dispatch is available for classes `carpet` and `cube` such that specific methods for these classes are called when given to the generic functions. p-values for the min-test are calculated for each combination group and multiplicity adjustment is done where needed. For global AVE- and MAX-tests, the approach proposed by Hung, Chi and Lipicki (1993) and Hung (2000) has been implemented as well as the resampling-based methods described in Chapters 4 and 5. Simultaneous confidence intervals can be calculated with an analogous syntax also based on the bootstrap or on the multivariate t-distribution if desired. The repeated calculations in all bootstrap algorithms are extremely computationally intensive and have therefore been implemented in `C++` for reasons of performance improvement.

As the R language offers powerful features of object oriented programming (OOP), it is apparent to represent the design and data from a clinical trial on combination drugs by an object with properties representing the dimension vector D, the measured data as a list of vectors for either treatment group and the sample size allocation n.
The syntax to construct objects of class `carpet` and `cube` is quite simple. For $k = 2$, it requires a list containing the data from the respective treatment groups ordered by rows, i.e. for a 2x2 design, the list should contain the data vectors in the order $(0,0)$, $(0,1)$, $(0,2)$, $(1,0)$, $(1,1)$, $(1,2)$, $(2,0)$, $(2,1)$ and finally $(2,2)$. To handle with the hypertension

example from Chapter 4, simulated data satifying the descriptive statistics given there can be generated by the code

```
library(bifactorial)
n<-c(75,75,74,48,74,75,74,49,48,50,48,48)
m<-c(0,1.4,2.7,4.6,1.8,2.8,5.7,8.2,2.8,4.5,7.2,10.9)
s<-rep(7.07,12)
x<-list(12)
for(i in 1:12){
  x[[i]]<-rnorm(n[i],mean=0,sd=1)
  x[[i]]<-x[[i]]-mean(x[[i]])
  x[[i]]<-x[[i]]*(s[i]/sd(x[[i]]))
  x[[i]]<-x[[i]]+m[i]
}
hung<-carpet(data=x,D=c(2,3))
```

where the latter constructs the `carpet` object `hung` from the simulated data. The generic functions `mintest`, `avetest`, `maxtest`, `confint` and S4 methods for the generic functions `plot`, `show` and `summary` from the `base` package are now available to be applied to the example. The constructed object is displayed by typing the object name `hung`:

```
Carpet size: 2 x 3

Sample size allocation matrix:
      0      1      2      3
0     75     75     74     48
1     74     75     74     49
2     48     50     48     48

Descriptive statistics: Mean response values
      0      1      2      3
0     0      1.4    2.7    4.6
1     1.8    2.8    5.7    8.2
2     2.8    4.5    7.2    10.9
```

```
Descriptive statistics: Standard deviations
      0      1      2      3
0     7.07   7.07   7.07   7.07
1     7.07   7.07   7.07   7.07
2     7.07   7.07   7.07   7.07
```

A graphical vizualisation can be generated by the command `plot(hung)`, the result of which was displayed in Figure 1.1a. Multiple inferences using the min-test and simultaneous confidence intervals can be conducted typing

```
mintest(hung,test="ttest",nboot=20000)
confint(hung,test="ttest",nboot=20000)
avetest(hung,test="ttest",nboot=20000)
maxtest(hung,test="ttest",nboot=20000)
```

The results from this have been presented in Chapter 4; the output looks like

```
Group     p-value
(1,1)     0.7280
(1,2)     0.0327
(1,3)     0.0398
(2,1)     0.5334
(2,2)     0.0096
(2,3)     2.0E-4
```

```
Contrast        Confidence interval
(1,1)-(1,0)     [-2.509;  4.455]
(1,1)-(0,1)     [-2.907;  4.844]
(1,2)-(1,0)     [ 0.379;  7.367]
(1,2)-(0,2)     [-0.521;  6.467]
(1,3)-(1,0)     [ 2.456; 10.284]
(1,3)-(0,3)     [-0.749;  7.883]
(2,1)-(2,0)     [-2.628;  5.961]
(2,1)-(0,1)     [-0.810;  6.950]
(2,2)-(2,0)     [ 0.028;  8.705]
(2,2)-(0,2)     [ 0.531;  8.408]
(2,3)-(2,0)     [ 3.728; 12.405]
(2,3)-(0,3)     [ 1.928; 10.605]
```

```
AVE-test for the existence of an efficacious combination
tave=2.4260, pave=0.0050

MAX-test for the existence of an efficacious combination
tmax=4.365, pmax<0.0001
Contrast with maximum test statistic: (2,3)
```

The multivariate-normal method proposed by Hung, Chi and Lipicky (1993) and Hung (2000) is available if the value `"hung"` is specified to the argument `method`, which has the default value `"bootstrap"` that was implicitly used above. The underlying code was implemented by Hellmich and Lehmacher in the context of their 2005 paper and is available at the URL `http://www.medizin.uni-koeln.de/kai/imsie/homepages/martin.hellmich/mfd.html`. For the confidence intervals, the calculations are based on the R package `multcomp` where algorithms by Genz, Bretz and Hothorn (2001) are used for numerical evaluation of multidimensional integrals for the multivariate t-distribution. The commands are

```
mintest(hung,test="ttest",method="hung")
confint(hung,test="ttest",method="hung")
avetest(hung,test="ttest",method="hung")
maxtest(hung,test="ttest",method="hung")
```

The output is given in the same format as above, but the particular results are not reproduced here as they were also given in Chapter 4. For binary data applications, a list of binary data, e.g. coded as 0 and 1 for "event" or "no event", respectively, can be specified to construct `carpet` objects. The value `"ztest"` can then be given to the argument `test` of the analysis methods such that e.g. the results for the AML example from Chapter 5 can be calculated. The graphical vizualization of this was displayed in Figure 1.1b. Note that no implementation of an analytical approach exists for this as nothing of this kind is available from literature.
All the methods except the graphical tools are also available for $k = 3$ designs. The source code of the package `bifactorial` can be downloaded from the CRAN network and it is distributed under the General Public License (GPL). The package vignette is attached on the following pages.

Package 'bifactorial' documentation

of

December 2, 2007

Title Inferences for bi- and trifactorial trial designs

Version 1.0

Date 2007-04-26

Depends R,methods,lattice,graphics,mvtnorm,multcomp

Author Peter Frommolt

Maintainer Peter Frommolt <peter.frommolt@medizin.uni-koeln.de>

Description This package makes global and multiple inferences for given bi- and trifactorial clinical trial designs using bootstrap methods and a classical approach.

License General Public License (GPL)

URL http://www.medizin.uni-koeln.de/kai/imsie/peter.frommolt/

R topics documented:

`avemax` *AVE- and MAX-test*

Description

Compute global tests for factorial dose-response designs following Hung (2000) or by a bootstrap algorithm.

Usage

```
avetest(C,test=NULL,method="bootstrap",nboot=NULL,simerror=NULL,
        output=TRUE,...)
maxtest(C,test=NULL,method="bootstrap",nboot=NULL,simerror=NULL,
        output=TRUE,...)
```

Arguments

`C`	An object of class `carpet` or `cube`.
`test`	Either `"ttest"` or `"ztest"` - the test statistic for the inferences to be based on. Use `"ztest"` for binary data applications.
`method`	The calculation method - use `"bootstrap"` for a resampling-based approach and `"hung"` for calculations using the multivariate normal distribution.
`nboot`	The number of bootstrap iterations to use.
`simerror`	Prespecified simulation standard error.
`output`	A logical variable indicating whether the calculation status should be displayed.
`...`	Any further arguments.

Details

When handling with data from factorial clinical trial designs, one is often interested in the question whether dose combinations in the trial have got a better effect than all of their component drugs, because regulatoric requirements demand a contribution to the efficacy by all components. The decision if any of the tested combination drugs has got this property can be based on the AVE- or MAX-statistics proposed by Hung, Chi and Lipicky (1993). The hypothesis that this is true for none of the combinations is rejected if the largest or the average of the min-statistics is sufficiently high. The functions `avetest` and `maxtest` calculate the corresponding p-values on `carpet` or `cube` objects with a new bootstrap algorithm, which is default, or by the multivariate method for unbalanced designs from Hung (2000). A resampling-based method is available also for binary data applications. The desired simulation accuracy always needs to be specified by the number `nboot` of simulations to perform or an upper bound `simerror` for the simulation standard error. If both are given, the two constraints will be held simultaneously. Depending on the type of data, the calculations can be based on Student's t-test for metric data or the Z-statistic for binary applications.

Value

An object of class `avetest` or `maxtext`, respectively, with the following slots. The slot `name` is available for the MAX-test only.

`p`	p-value for the AVE- or MAX-test.
`stat`	Observed AVE- or MAX-statistic.
`test`	Type of test statistic which the AVE- or MAX-test was based on.
`method`	Algorithm used for the calculation.
`nboot`	Total number of resampling iterations.
`simerror`	Simulation standard error.
`name`	Combination group where the maximum of the min-statistics was observed.
`duration`	Total computing duration in seconds.
`call`	The function call.

Note

The performance of the bootstrap-based approach and the method from Hung (2000) has been compared and discussed. All algorithms perform very conservative if the means in the marginal treatment groups are close for the combinations.

Author(s)

Peter Frommolt, University of Cologne ⟨peter.frommolt@medizin.uni-koeln.de⟩
`http://www.medizin.uni-koeln.de/kai/imsie/peter.frommolt/`

References

Hellmich M, Lehmacher W (2005): Closure procedures for monotone bi-factorial dose-response designs. Biometrics 61, pp. 269-276

Hung HMJ, Chi GYH, Lipicky RJ (1993): Testing for the existence of a desirable dose combination. Biometrics 49, pp. 85-94

Hung HMJ, Wang SJ (1997): Large-sample tests for binary outcomes in fixed-dose combination drug studies. Biometrics 53, pp. 498-503

Hung HMJ (2000): Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. Statistics in Medicine 19, pp. 2079-2087

See Also

`bifactorial`, `carpet`, `cube`, `mintest`, `margint`

Examples

```
#Hypertension example from Hung (2000)
n<-c(75,75,74,48,74,75,74,49,48,50,48,48)
m<-c(0,1.4,2.7,4.6,1.8,2.8,5.7,8.2,2.8,4.5,7.2,10.9)
s<-rep(7.07,12)
x<-list(12)
for(i in 1:12){
  x[[i]]<-rnorm(n[i],mean=0,sd=1)
  x[[i]]<-((x[[i]]-mean(x[[i]]))*(s[i]/sd(x[[i]])))+m[i]
}
hung<-carpet(x,D=c(2,3))
avetest(hung,test="ttest",nboot=20000)
maxtest(hung,test="ttest",nboot=20000)
```

`bifactorial` *General information on the package*

Description

Factorial clinical trial designs can be used to test for the efficacy of combination drugs with two or more components, where inference on the question if a combination therapy is more efficacious than both of its components is based on the min-test proposed by Laska and Meisner (1989). This is due to regulatoric demands requiring a contribution of all compounds in a combination drug. The AVE- and MAX-approaches proposed by Hung, Chi and Lipicky (1993) test for the existence of any desirable combination.

Bootstrap-based methods are implemented as well as classical approaches available from literature to obtain p-values and confidence intervals in such designs. For the min-test, analytical methods use a normality and homoscedasticity assumption on the data (Hung, Chi and Lipicky, 1993 and Hung, 2000). Critical values needed for determination of confidence intervals are calculated using quantiles of the multivariate t-distribution (Bretz, Genz and Hothorn 2001). These methods fail when handling with data that are skewed or heteroscedastic over the treatment groups. Furthermore, no analytical approach is available for the trifactorial case and the AVE- and MAX-tests on binary data. In the bootstrap approach, only the empirical distribution of the data is used and thus the results are valid for any distributional shape, provided that sufficiently large samples are available. Less analytical framework is needed to handle with the distributional properties of the tests. Further information on resampling-based methods and theoretical backgrounds are given in Westfall and Young (1993).

Anyway, the problem of the extremely decreasing power for small values of the so-called nuisance parameters indicating the response differences between the marginal treatment groups cannot be resolved by the bootstrap approach. Any algorithm based on estimates for the nuisance parameters other than the assumption that they are infinite will exceed the given type I error level (Snapinn, 1987).

The package contains the generic functions `mintest` and `margint` to test for mean differences of given numeric data vectors and differences in event rates for binary data applications. Method dispatch is available for objects of class `carpet` or `cube`, which will lead to min-test results on a bi- or trifactorial design and corresponding confidence intervals comparing

combination treatments with their respective component therapies. Implementations for global tests are also available by the generic functions `avetest` and `maxtest`.

Author(s)

Peter Frommolt, University of Cologne ⟨peter.frommolt@medizin.uni-koeln.de⟩
`http://www.medizin.uni-koeln.de/kai/imsie/homepages/peter.frommolt/`

References

Bretz F, Genz A, Hothorn LA (2001): On the numerical availability of multiple comparison procedures. Biometrical Journal 43/5, pp. 645-656

Hellmich M, Lehmacher W (2005): Closure procedures for monotone bi-factorial dose-response designs. Biometrics 61, pp. 269-276

Hung HMJ, Chi GYH, Lipicky RJ (1993): Testing for the existence of a desirable dose combination. Biometrics 49, pp. 85-94

Hung HMJ, Wang SJ (1997): Large-sample tests for binary outcomes in fixed-dose combination drug studies. Biometrics 53, pp. 498-503

Hung HMJ (2000): Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. Statistics in Medicine 19, pp. 2079-2087

Laska EM, Meisner MJ (1989): Testing whether an identified treatment is best. Biometrics 45, pp. 1139-1151

Snapinn SM (1987): Evaluating the efficacy of a combination therapy. Statistics in Medicine 6, pp. 657-665

Westfall PH, Young SS (1993): Resampling-based multiple testing. John Wiley & Sons, Inc., New York

---

`carpetcube` *Objects for handling with bi- and trifactorial trial data*

Description

Create objects representing bi- or trifactorial clinical trial designs.

Usage

```
carpet(data,D,...)
cube(data,D,...)
```

Arguments

`data`	A list of numeric or binary data vectors from the trial. See the details below for the order in which the list is to be given.
`D`	An integer vector of length 2 for `carpet` objects and of length 3 for `cube` objects, specifying the number of doses of the components drugs in the trial.
`...`	Any further arguments.

Details

The function `carpet` creates objects of class `carpet` from the specified `data` in the list that are used row-wise to fill up the 2-factorial treatment groups, i.e. in the order (0,0), (0,1),..., (0,D[2]), (1,0), ..., (1,D[2]), ..., (D[1],D[2]); resulting in a (D[1]+1)x(D[2]+1) data array.

To represent trifactorial designs for the evaluation of a three-compound combination, an object of class `cube` can be created using the function `cube`. The data in the treatment groups are then filled up in the order (0,0,0), ..., (0,0,D[3]) first, then (0,1,0), ..., (0,1,D[3]) and up to (0,D[2],0), ..., (0,D[2],D[3]). This is the order also for the values 0, ..., D[1] for the first component group, always taking the data succesively from the list elements of `data`. The result is a (D[1]+1)x(D[2]+1)x(D[3]+1) data array. Methods for multiple inference and global tests can be applied to `carpet` and `cube` objects.

Value

An object of class `carpet` or `cube`, respectively, with the following slots.

`data`	The data list specified in the construction.
`D`	Vector of maximum doses specified in the construction.
`n`	Numeric vector of sample sizes in the respective groups.

Author(s)

Peter Frommolt, University of Cologne ⟨peter.frommolt@medizin.uni-koeln.de⟩
`http://www.medizin.uni-koeln.de/kai/imsie/homepages/peter.frommolt/`

References

Hung HMJ, Chi GYH, Lipicky RJ (1993): Testing for the existence of a desirable dose combination. Biometrics 49, pp. 85-94

Hung HMJ (2000): Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. Statistics in Medicine 19, pp. 2079-2087

See Also

`bifactorial`, `mintest`, `margint`, `avetest`, `maxtest`

Examples

```
#Hypertension example from Hung (2000)
data(sidbp)
x<-split(sidbp$ynrmhom,sidbp$cb)
bifactorial<-carpet(data=x,D=c(2,3))
```

`multinf` *Multiple inference*

Description

Compute adjusted p-values and simultaneous confidence intervals for given bi- and trifactorial design data.

Usage

```
mintest(C,test=NULL,method="bootstrap",nboot=NULL,simerror=NULL,
        output=TRUE,...)
margint(C,test=NULL,method="bootstrap",nboot=NULL,simerror=NULL,
        alpha=0.05,output=TRUE,...)
```

Arguments

`C`	An object of class `carpet` or `cube`.
`method`	The calculation method - use `"bootstrap"` for a resampling-based approach, `"hung"` for the min-test approach of Hung (2000) and `"tdistr"` for interval calculations based on the multivariate t-distribution.
`test`	Either `"ttest"` or `"ztest"` - the test statistic for the inferences to be based on. Use `"ztest"` for binary data applications.
`alpha`	Simultaneous level of the confidence intervals.
`nboot`	Number of resampling iterations to use.
`simerror`	Prespecified simulation standard error.
`output`	Logical value to specify whether a status output should be given during calculation.
`...`	Any further arguments.

Details

The generic functions `mintest` and `margint` calculate adjusted p-values and simultaneous confidence intervals for the test of parametric differences between prespecified treatment groups on bi- or trifactorial design clinical trials. If an object of class `carpet` is commited, `mintest` will return adjusted p-values for the min-test on combination superiority in bifactorial clinical trial designs (Laska and Meisner, 1989). The alternative hypothesis of this test is that the detected effect size for the combination treatment is better than for both single component groups; i.e. the test results in only one p-value for each combination. The generic function `margint` will, when applied to `carpet` objects, return simultaneous confidence intervals for the parametric differences between each combination treatment group and its respective components. Depending on the type of data, the calculations can be based on Student's t-test for metric data or the Z-statistic for binary applications.

By default, the calculations are performed by a resampling-based approach. The desired simulation accuracy always needs to be specified by the number `nboot` of bootstrap iterations to perform or an upper bound `simerror` for the simulation standard error. If both are

given, the two constraints will be held simultaneously. On the other hand, the multivariate normal approach for unbalanced designs from Hung (2000) is available when the argument `method` is set to the value `"hung"`. For the trifactorial case, no such approach is available and thus the calculations are based on the bootstrap approach, performing a generalized min-test on the data, if an object of class `cube` is commited. The interval calculations are based on the multivariate t-distribution if `"tdistr"` is specified.

In the classical approach to the min-test, a normality assumption for the data is used and the desired critical values are calculated using quantiles of the multivariate t-distribution. However, this method fails when handling with data that are skewed or heteroscedastic over the treatment groups. When using the bootstrap, only the empirical distribution of the data is used and thus the results are always valid, provided that a sufficiently large samples are available. When handling with data from bifactorial clinical trial designs, bootstrap methods need much less analytical framework on the distributional properties of the tests than if the approach given by Hung (2000) is used. In particular, the restriction to only two compounds is not needed and binary data applications can be handled analogously. The theory of resampling-based multiple testing has been extensively discussed by Westfall and Young (1993).

The calculation of simultaneous confidence intervals is much easier because the c.d.f. of the min-statistic is not needed. Hence this is leading to an ordinary multiple contrast problem.

Value

An object of class `mintest` or `margint` with the following slots.

`p`	Adjusted p-values for the respective combination groups.
`stat`	The observed values of the min-statistics.
`kiu`	The lower limits of the confidence intervals.
`kio`	The upper limits of the confidence intervals.
`alpha`	One minus the nominal coverage probability of the confidence intervals.
`gnames`	Names of the combination groups.
`cnames`	The names of the contrasts for comparisons of the combinations with their respective components.
`test`	Type of test statistics that the min-tests were based on.
`method`	The method used for calculation.
`nboot`	Number of bootstrap replications used.
`simerror`	Maximum of the simulation standard errors in the combination groups.
`duration`	Total computing duration in seconds.
`call`	Function call.

Note

Performance of the implemented methods has been evaluated and compared. The min-test performs very conservative if the means in the marginal treatment groups are close for a combination.

Author(s)

Peter Frommolt, University of Cologne ⟨peter.frommolt@medizin.uni-koeln.de⟩
http://www.medizin.uni-koeln.de/kai/imsie/homepages/peter.frommolt/

References

Hung HMJ, Chi GYH, Lipicky RJ (1993): Testing for the existence of a desirable dose combination. Biometrics 49, pp. 85-94

Hung HMJ, Wang SJ (1997): Large-sample tests for binary outcomes in fixed-dose combination drug studies. Biometrics 53, pp. 498-503

Hung HMJ (2000): Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. Statistics in Medicine 19, pp. 2079-2087

Hellmich M, Lehmacher W (2005): Closure procedures for monotone bi-factorial dose-response designs. Biometrics 61, pp. 269-276

Laska EM, Meisner MJ (1989): Testing whether an identified treatment is best. Biometrics 45, pp. 1139-1151

Snapinn SM (1987): Evaluating the efficacy of a combination therapy. Statistics in Medicine 6, pp. 657-665

Westfall PH, Young SS (1993): Resampling-based multiple testing. John Wiley & Sons, Inc., New York

See Also

`bifactorial`, `carpet`, `cube`, `avetest`, `maxtest`,

Examples

```
#AML example from Huang et al. (2007) with data from
#Issa et al. (2004) and Petersdorf et al. (2007)
n<-c(10,31,17,100,50,50,101,50,50)
p<-c(0.00,0.45,0.65,0.30,0.71,0.70,0.59,0.64,0.75)
y<-list()
for(i in 1:9){
  y[[i]]<-0
  while((sum(y[[i]])!=round(n[i]*p[i]))||(length(y[[i]])==1)){
    y[[i]]<-rbinom(n[i],1,p[i])
  }
}

aml<-carpet(data=y,D=c(2,2))
mintest(aml,test="ztest",nboot=25000)
margint(aml,test="ztest",nboot=25000)
```

`sidbp` *Data of sitting distolic blood pressure (SiDBP)*

Description

These data have been simulated with respect to the descriptive statistics given in the bifactorial hypertension clinical trial reported by Hung (2000). Various distributional properties have been realized for normal and skewed cases with equal variances as well as with linearly increasing variances. The latter means that the coefficient of variation is held constant over the treatment groups. The group defining variable is named `cb` in the data set. It has got the levels `(0,0)`, `(1,0)`, `...`, `(2,3)` according to the respective dose combinations. The data vectors with the different distributional properties are `ynrmhom`, `ynrmhet`, `yloghom` and `yloghet`.

Format

A data frame with 738 observations on the following 5 variables.

`cb` a factor with levels `(0,0)`, `(0,1)`, `...`, `(2,3)`

`ynrmhom` A vector of normal and homoscedastic data.

`ynrmhet` A vector of normal and heteroscedastic data.

`yloghom` A vector of lognormal and homoscedastic data.

`yloghet` A vector of lognormal and heteroscedastic data.

References

Hung HMJ (2000): Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. Statistics in Medicine 19, pp. 2079-2087

Examples

```
data(sidbp)
```

Mein Lebenslauf wird aus Gründen des Datenschutzes in der elektronischen Fassung meiner Arbeit nicht veröffentlicht.