The Likelihood Ratio Test for Order Restricted Hypotheses in Non-Inferiority Trials

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von

Guido Skipka
aus Simmerath

Göttingen, den 28.5.2003

Referent: Prof. Dr. A. Munk
Korreferent: Prof. Dr. M. Denker
Tag der mündlichen Prüfung: 25.6.2003

Contents

1 Introduction

In clinical investigations with the goal to evaluate new therapies or diagnostic methods, the randomized controlled trial is the established design of a study. Especially in drug development the randomized controlled trial design is mandatory prescribed by regulatory agencies. In the past decades the efficacy of a therapy was declared by its superiority over a placebo. Nowadays, for many indications and diseases the clinical progress leads to the ethical problem of choosing placebo as a control. The declaration of Helsinki which expresses the ethical principles for medical research involving human subjects of the World Medical Association, states: "*The benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where no proven prophylactic, diagnostic or therapeutic method exists*". Therefore, today it is common to compare a new therapy with an established standard therapy (so-called *active control*). This implies smaller differences between a new therapy and its control resulting in a very large number of patients to achieve a given power for detecting a difference between the groups. For acute myocardial infarction trials to evaluate thrombolytic agents, more than 40,000 patients had to be enrolled in a study (GUSTO [1993]). Therefore, in the last decade new approaches became popular focussing on the equivalence of a new therapy and an established standard and not on superiority, see e.g. Dunnett and Gent [1977], Blackwelder [1982], Farrington and Manning [1990], Kieser [1995], Chan [1998], Röhmel and Mansmann [1999b], Kang and Chen [2000], Hoover and Blackwelder [2001], Phillips [2003], Pigeot et al. [2003]. With this approach the sample sizes can be reduced substantially. Further motivations for equivalence trials are less side effects, less adverse events or an easier handling/application of the new therapy.

In clinical trials comparing a new therapy and a standard, the investigators mostly want to show that the new therapy is superior to its comparator. This leads to one-sided hypotheses. In almost all cases the term equivalence in the manner described above is also defined by an one-sided hypothesis since, more precisely, the goal is to show that the new therapy is at most irrelevantly inferior to the standard and may be much better. Equivalence in this context is often called *therapeutic equivalence* in the literature. However, the term equivalence suggests a restriction to both directions. It is also used in bioequivalence trials which involve two-sided hypotheses. For this reason, in the following

the term *equivalence* will be avoided and replaced by the term *non-inferiority*.

The parallel group design is the most common setting in clinical trials evaluating new therapies or diagnostics. The patients are randomly allocated to two or more groups with different treatment strategies. For quantifying a therapeutic effect θ, usually measures of location are specified. For continuous outcomes the difference, the standardized difference or the ratio of the endpoints are mostly used. For binary outcomes the difference, the ratio (in medicine often called risk ratio or relative risk) or the odds ratio of the rates is specified.

If a non-inferiority trial is planned, the term *inferior* has to be quantified by defining the equivalence margin. This margin is the largest difference which is still clinically acceptable. A difference bigger than this margin would matter. There is a broad still ongoing debate how to specify an equivalence margin. A general rule cannot be given. The equivalence margin depends on clinical aspects like the indication or the endpoint. Furthermore, the margin is oriented on results of earlier clinical trials. The question how to specify this margin will not be pursued here.

It is a structural problem to ensure the external validity in two-armed non-inferiority trials comparing a new treatment and a standard therapy. The external validity is the ability of a trial to allow for the conclusion of the efficacy of the new treatment, if non-inferiority is shown. If the efficacy of the standard (the difference between standard and placebo, where placebo is not evaluated in the trial) is smaller than the specified equivalence margin, then non-inferiority of an inefficacious new treatment is shown. The ability of a trial to evaluate the efficacy of a new treatment is called *assay sensitivity*. For two-armed trials without a placebo group this can be intended only by consulting historical comparisons, i.e. investigating the results of similar trials in the past (so-called *meta analyses*). If the information about the efficacy of the standard treatment is insufficient, a third placebo group may be justifiable. In these cases it is possible to prove the assay sensitivity directly. Therefore, statistical tests are investigated for two and three independent groups.

Closely related to this is the aim to establish a relevant superiority of a new treatment compared to a standard one in superiority trials, see e.g. Chan [1998], Chuang-Stein [2001], Dunnett and Tamhane [1997], Greco et al. [1996], Gustafsson et al. [1996], Moulton et al. [2001], Röhmel and Mansmann [1999b]. The most common ways to deal with this problem is to test a proper hypothesis (to be described later on) or to base the decision on a confidence interval (see e.g. Newcombe [1998] for a survey).

For comparing two groups, a discrepancy measure is defined which is used to describe the aim of the clinical trial. In the following it is assumed that θ is the measure for the inferiority of the new treatment T in comparison to the control C. Choosing θ as the

difference or standardized difference, the strict equivalence of both groups is expressed by $\theta = 0$. The statistical hypotheses for proving non-inferiority of T in comparison to C are given by

$$H_0 : \theta \geq \theta_0 \quad \text{vs.} \quad H_1 : \theta < \theta_0 \ , \tag{1.1}$$

where the equivalence margin θ_0 has to be specified larger than 0.

On the other hand, if the goal is to prove the relevant superiority of T, the margin θ_0 in (1.1) has to be chosen smaller than 0.

For the relative risk or the odds ratio the strict equivalence holds for $\theta = 1$. Here the margin θ_0 has to be chosen > 1 for non-inferiority and < 1 for relevant superiority, respectively.

Examples: The efficacy of a new thrombolytic agent (T) in comparison to a standard one (C) was investigated in patients with acute myocardial infarction by Tebbe et al. [1998]. Non-inferiority of T over C was defined in terms of the odds ratio of the 30-days mortality rates, with an equivalence margin $\theta_0 = 1.5$. The hypotheses were

$$H_0 : \frac{p_T(1-p_C)}{p_C(1-p_T)} \geq 1.5 \quad \text{vs.} \quad H_1 : \frac{p_T(1-p_C)}{p_C(1-p_T)} < 1.5 \ ,$$

where p_T and p_C represented the mortality rates of the two groups.

In a three-armed trial by Diehm et al. [1996] in patients with chronic venous insufficiency the efficacy of dried horse chestnut seed extract was examined. Primary endpoint was the reduction of the lower leg volume after a treatment over a period of 12 weeks. Two control groups were included: compression stockings (C) and a drug placebo (P). Based on the standardized difference of means two goals were intended: first, the relevant superiority of C as compared to P, second, the non-inferiority of the new treatment (T) as compared to C. An equivalence margin of $\theta_0 = 0.5$ was chosen for both comparisons. Therefore, the hypotheses

$$\begin{aligned} H_0 : \frac{\mu_P - \mu_C}{\sigma} \geq 0.5 \quad \text{vs.} \quad H_1 : \frac{\mu_P - \mu_C}{\sigma} < 0.5 \quad & \text{(rel. superiority)}, \\ H_0 : \frac{\mu_C - \mu_T}{\sigma} \geq -0.5 \quad \text{vs.} \quad H_1 : \frac{\mu_C - \mu_T}{\sigma} < -0.5 \quad & \text{(non-inferiority)}, \end{aligned}$$

where specified with μ_T, μ_C, μ_P as the means and σ as the standard deviation of the groups, respectively. Since the effect of the compression stockings was not well known at the planning phase, the trial was evaluated by means of a hierarchical test procedure: Non-inferiority of the new therapy with respect to the standard should be shown in a second step, provided the relevant superiority of the standard over placebo was shown in a first step.

For normally distributed data Pigeot et al. [2003] suggest a three-armed design for a treatment group (T), an active control group (C) and a placebo group (P). They specify the null hypothesis

$$H_0 \colon \mu_C - \mu_T \geq p\,(\mu_P - \mu_C)\ , \quad -1 < p < 0$$

to avoid the problem of assay sensitivity. Using this hypothesis, the equivalence margin is specified as a fraction of the efficacy of the control (difference between C and P).

In this work, hypotheses for three-armed designs consisting of two pairwise hypotheses of the type (1.1) are considered.

As before, one or two new treatment groups are denoted by T (T_1, T_2, respectively), one or two active control groups as C (C_1, C_2, respectively), and a placebo group as P.

From a medical point of view the following problems including non-inferiority for three-armed trials may be of interest:

1. *non-inferiority of T w.r.t. S_1 and w.r.t. S_2*: In this situation the new treatment has to be at least as effective as two standards.
2. *non-inferiority of T w.r.t. S_1 or w.r.t. S_2*: In this situation the new treatment has to be at least as effective as one of the two or both standards.
3. *non-inferiority of T_1 or T_2 w.r.t. S*: In this situation one of the two or both new treatments have to be at least as effective as the standard.
4. *non-inferiority of T_1 and T_2 w.r.t. S*: In this situation both new treatments have to be at least as effective as the standard, respectively.
5. *non-inferiority of T w.r.t. S and superiority of S w.r.t. P*: In this situation the new treatment has to be at least as effective as the standard and, simultaneously, the standard has to be relevantly more effective than the placebo.
6. *non-inferiority of T w.r.t. S and superiority of T w.r.t. P*: In this situation the new treatment has to be at least as effective as the standard and, simultaneously, relevantly more effective than the placebo.

From a statistical point of view the problems mentioned above can be expressed by the following three types of hypotheses. Let $\theta_{i,j}$ denote the distance measure for group i and j, and let $\theta_{0_1}, \theta_{0_2}$ be the two equivalence margins, respectively:

a) $H_0 \colon \theta_{1,2} \geq \theta_{0_1} \ \vee \ \theta_{1,3} \geq \theta_{0_2}$ vs. $H_1 \colon \theta_{1,2} < \theta_{0_1} \ \wedge \ \theta_{1,3} < \theta_{0_2}$,
b) $H_0 \colon \theta_{1,2} \geq \theta_{0_1} \ \wedge \ \theta_{1,3} \geq \theta_{0_2}$ vs. $H_1 \colon \theta_{1,2} < \theta_{0_1} \ \vee \ \theta_{1,3} < \theta_{0_2}$,
c) $H_0 \colon \theta_{1,2} \geq \theta_{0_1} \ \vee \ \theta_{2,3} \geq \theta_{0_2}$ vs. $H_1 \colon \theta_{1,2} < \theta_{0_1} \ \wedge \ \theta_{2,3} < \theta_{0_2}$.

The problems 1, 4 and 5 are of type a), the problems 2 and 3 are of type b) and the problem 6 is of type c).

The goal of this work is to derive statistical tests based on the likelihood ratio (LR) principle for the three hypotheses mentioned above. Further we will compare these tests with the commonly used statistical approaches with respect to level and power. We will focus solely on testing methods. However, in principle all procedures can be used to obtain confidence intervals by proper inversion (Casella and Berger [2002, Ch. 9.2]).

This work is organized as follows: In Chapter 2 we introduce the general methodology of LR tests. The main theorems will be given in order to determine the asymptotic distribution of the likelihood ratio statistic for special cases introduced in the subsequent chapters. In Chapter 3 we give a survey for the two-sample case for normally distributed data. It will be shown that the LR tests for the commonly used distance measures are equivalent to the t-tests usually applied in this situation. This will be extended to three groups in Chapter 4. The LR test will be derived applying the methodology of *order restricted inference* (see e.g. Robertson et al. [1988]). We show that for the hypotheses a) and c), respectively, the LR test is the same as the intersection-union test, when its two-sample pooled variances are replaced by the three-sample pooled variance. Robertson et al. [1988] derived the LR test in the general setting of $k > 2$ groups which will be used for the hypotheses b). In this case it is possible to give explicit formulae under the assumption of three homoscedastic groups. We investigate the power of the LR test and the commonly used pairwise comparison procedures. It will be found that the LR test is comparable and sometimes slightly superior to the best of the competitors.

In Chapter 5 the LR test for binary outcomes is investigated in the two-sample situation. We derive the asymptotic distribution of the likelihood ratio statistic for general distance measures in Section 5.2. A comparison to other asymptotic approaches will be given regarding the commonly used distance measures. We will show that the power differences are only marginal on the one hand. On the other hand, the LR test keeps the level more accurately than its competitors.

It is well known from the literature that the actual level of asymptotic approaches exceeds the nominal level for small sample sizes. In Section 5.3 we suggest an exact version of the LR test which is based on an idea used by Storer and Kim [1990] in a different context. This test is analyzed and compared in a numerical study with various competitors from the literature. It will be found that in general the power of the LR test tends to be larger, even if the improvement is small. On the other hand, the computational effort is larger for the LR test. As a by-product which is of interest on its own we observe serious numerical difficulties with Barnard's [1947] test. We will analyze these difficulties and give an explanation for them.

Additionally, we will discuss briefly sample size calculations, and we will show that the power is maximized in general for unequal group sample sizes in non-inferiority trials. This is in contrast to trials where the null hypothesis states the equality of treatments.

In Chapter 6 the asymptotic and exact LR approaches are extended to three groups for the binomial distribution. We will derive the asymptotic distribution of the LR test for the hypotheses mentioned above. Analogously to the case of normally distributed data, we obtain the pairwise comparison procedure for hypotheses a) and c). For the hypotheses b), the LR approach is different from the pairwise comparisons and its asymptotic distribution can be determined analytically. The asymptotic distribution, however, depends on unknown nuisance parameters and cannot be used for the practical performance of the test. Therefore, we suggest to substitute the constrained maximum likelihood estimator for the unknown parameters and to determine the distribution numerically. In a numerical study we show that this approach is comparable to other asymptotic pairwise procedures with respect to level and power. Analogously to the two-sample case, we construct an unconditional exact version based on the LR statistic. This test will be extensively compared to the pairwise exact procedures for various distance measures. We find that the exact LR test represents an improvement on the corresponding pairwise two-sample procedures which can be quite substantial in various cases.

In summary, the exact version of the LR statistic represents a unified and powerful tool for the assessment of non-inferiority in two- and three-sample settings. In particular, for binary responses the improvement in power can become quite substantial compared to its competitors from the literature.

Acknowledgements

I am very grateful to my supervisor Professor Dr. A. Munk for proposing this subject and for many helpful discussions. I also would like to thank Prof. Dr. H. J. Trampisch, Dr. N. Bissantz, Dr. G. Freitag, Dr. S. Koch and Dr. B. Stratmann for useful discussions. My special thanks go to Dr. S. Lange for many discussions and comments regarding the medical and biometrical topics and his steady readiness to help.

2 The likelihood ratio principle

The likelihood ratio (LR) principle is a general parametric method to derive statistical tests for parameters of a probability distribution. Especially for composite hypotheses which are hypotheses including more than one distribution, this principle leads to very powerful tests, in general.

In this chapter the general methodology for likelihood ratio tests is described for independent samples. This will be used in the subsequent sections to derive the likelihood ratio test for specific designs and settings.

Let stochastically independent random variables $X_1, \ldots, X_n$ be given, where X_i has a density function f_i, $i = 1, \ldots, n$, depending on a parameter vector $\vartheta \in \Theta \subseteq \mathbb{R}^k$. For a fixed sample $x_1, \ldots, x_n$ the *likelihood function*

$$L(\vartheta) := \prod_{i=1}^{n} f_i(x_i, \vartheta)$$

provides the probability for obtaining the sample as a function of ϑ, if the sample space is discrete. For continuous outcomes the likelihood function provides a value which is proportional to the probability that a sample lies in the neighborhood of the observed sample. If the null hypothesis and alternative hypothesis divide the parameter space $\Theta = \Theta_0 \cup \Theta_1$ in two disjoint sets, the *likelihood ratio* (LR) is given by

$$\lambda := \frac{L(\hat{\vartheta}^*)}{L(\hat{\vartheta}\,)} = \frac{\sup_{\vartheta \in \Theta_0} L(\vartheta)}{\sup_{\vartheta \in \Theta} L(\vartheta)} \; . \tag{2.1}$$

Any parameter vector maximizing the density is called a *maximum likelihood estimator* (MLE). Therefore, the LR is the ratio of the joint density with the MLE constrained to the null hypothesis, $\hat{\vartheta}^*$, and of the density with the unconstrained MLE $\hat{\vartheta}$. If the unknown true parameter vector is included in the null hypothesis, the LR tends to 1, otherwise it tends to 0. Thus, it is tempting to use the LR as a test statistic for constructing a statistical test. For many models, under certain requirements the random variable $-2 \log \lambda$ follows a χ^2-law.

If the null space Θ_0 is multidimensional, the numerical determination of the LR is very cumbersome. The following Theorem 2.1 provides conditions under which the maximum

is located at the boundary $\partial\Theta_0$ of Θ_0. This makes the determination of the maximum much easier.

Theorem 2.1 *Let $\vartheta \in \Theta \subseteq \mathbb{R}^k$ and $\Theta_0 \subset \Theta$, where Θ_0 and Θ are closed. Let $X_1, \ldots, X_n$ be a vector of independent random variables with densities f_i, respectively, such that for each realization $x_1, \ldots, x_n$*

$$\lim_{\|\vartheta\| \to \infty} f_i(x_i, \vartheta) = 0 \tag{2.2}$$

holds. Let the likelihood function $L(\vartheta) = \prod_{i=1}^n f_i(x_i, \vartheta)$ be differentiable in ϑ and grad $L(\vartheta) \neq 0$ for all $\vartheta \in \overset{\circ}{\Theta}_0 := \Theta_0 \setminus \partial\Theta_0$.

Then, the constrained MLE exists, and we have for any MLE

$$\{\tilde{\vartheta} \mid \tilde{\vartheta} := \arg\max_{\vartheta \in \Theta_0} L(\vartheta)\} \subseteq \partial\Theta_0 \ .$$

Proof: In a first step we show the existence of a maximizer of L, i.e. the maximum is attained in Θ. For compact Θ_0 the assertion is trivial. Therefore, assume that Θ_0 is not compact. The compact sets $K_m := \{\vartheta \in \Theta_0 | \ \|\vartheta\| \leq m\}$ $(m \in \mathbb{N})$ cover Θ_0, i.e. $\bigcup_{m=1}^{\infty} K_m = \Theta_0$. There is a maximizer ϑ_m of L restricted to K_m, again due to compactness of K_m. The sequence K_m being increasing, i.e. $K_m \subseteq K_{m+1}$, implies $L(\vartheta_m)$ to be increasing. Suppose there is a subsequence ϑ_{m_l} such that $\|\vartheta_{m_l}\| \to \infty$. The Assumption (2.2) implies $\lim_{\|\vartheta\| \to \infty} L(\vartheta) = 0$, in particular $\lim_{l \to \infty} L(\vartheta_{m_l}) = 0$. Since $L \not\equiv 0$ and $L \geq 0$, this contradicts the monotonicity of $L(\vartheta_{m_l})$. Hence, $\|\vartheta_m\|$ is bounded, i.e. there is a compact set $K \subset \Theta_0$ such that $\vartheta_m \in K$ for all m. Thus, maximization is restricted to this compact set K which proves the assertion.

Now it is shown that the maximum is attained at the boundary of Θ_0. Suppose the contrary, i.e. $\tilde{\vartheta} \in \overset{\circ}{\Theta}_0$. Then there exists an open neighborhood $U_0 \subseteq \overset{\circ}{\Theta}_0$ of $\tilde{\vartheta}$, so that $\tilde{\vartheta}$ is a local maximum in U_0. Since $L(\vartheta)$ is differentiable for all $\vartheta \in \Theta_0$, it follows that grad$L(\tilde{\vartheta}) = 0$ which is a contradiction. □

The merit of Theorem 2.1 is twofold. First, it allows to restrict the parameter space to a one dimensional curve Θ_h for numerical computation of the constrained MLE, if $\Theta \subseteq \mathbb{R}^2$. Second, this will be the key property for the derivation of the asymptotic distribution of the likelihood ratio statistic λ.

The following results from Pruscha [2000] are required to obtain the asymptotic distribution of the LR statistic for different models and hypotheses.

Let

$$U_n(\vartheta) := \frac{\partial}{\partial\vartheta} \log L(\vartheta)$$

denote the k-dimensional score vector of the likelihood function and

$$W_n(\vartheta) := \frac{\partial}{\partial \vartheta} U_n^\top(\vartheta)$$

the $(k \times k)$-functional matrix of $U_n(\vartheta)$.

Definition 2.2 *Let Γ_n be a $(k \times k)$-diagonal matrix with positive elements and $\Gamma_n \to 0$ for $n \to \infty$. A sequence $\hat{\vartheta}_n$ of k-dimensional random vectors is called a Γ_n^{-1}-consistent Z-estimator for ϑ, if for all $\vartheta \in \Theta$,*

1.) $P_\vartheta(U_n(\hat{\vartheta}_n) = 0) \longrightarrow 1 \quad (n \to \infty)$,

2.) $\Gamma_n^{-1}(\hat{\vartheta}_n - \vartheta)$ *is P_ϑ-stochastically bounded,*
i.e. $\lim_{M\to\infty} \limsup_{n\to\infty} P_\vartheta(|\Gamma_n^{-1}(\hat{\vartheta}_n - \vartheta)| > M) = 0$.

The following Theorem provides conditions for the asymptotic normality of a Z-estimator.

Theorem 2.3 *Let $\Sigma(\vartheta)$ and $B(\vartheta)$ denote positive definite $(k \times k)$-matrices. If for all $\vartheta \in \Theta$ and for $n \to \infty$ the conditions*

1.) $\Gamma_n U_n(\vartheta) \xrightarrow{\mathcal{D}} N_k(0, \Sigma(\vartheta))$,

2.) $\Gamma_n W_n(\vartheta_n^*)\Gamma_n \xrightarrow{P_\vartheta} -B(\vartheta)$ *for all sequences of random vectors ϑ_n^*, for which $\Gamma_n^{-1}(\vartheta_n^* - \vartheta)$ is P_ϑ-stochastically bounded,*

hold, it follows for a Γ_n^{-1}-consistent Z-estimator $\hat{\vartheta}_n$ for ϑ that

$$\Gamma_n^{-1}(\hat{\vartheta}_n - \vartheta) \xrightarrow{\mathcal{D}} N_k(0, (B^{-1}(\vartheta))^\top \Sigma(\vartheta) B^{-1}(\vartheta)) ,$$

where $N_k(\mu, \Sigma)$ denotes the k-dimensional normal distribution with mean μ and covariance matrix Σ.

Proof: Pruscha [2000, p. 194] □

The parametrization of composite hypotheses is given by a $C^{(2)}$-function
$h : \Delta \to \Theta$, $h(\eta) = (h_1(\eta), \ldots, h_k(\eta))^\top$, $\eta \in \Delta$, over a parameter subspace $\Delta \subset \mathbb{R}^c$ with $1 \le c < k$. Let

$$h'(\eta) := (\frac{\partial}{\partial \eta} h(\eta))^\top ,$$
$$h_j''(\eta) := \frac{\partial^2}{\partial \eta \partial \eta^\top} h_j'(\eta) \ , \ j = 1, \ldots, k \ ,$$

be the functional $(k \times c)$-matrix of h with full rank c and the Hessian $(c \times c)$-matrix of h_j, $j = 1, \ldots, k$, respectively.

Theorem 2.4 *Let $C(\eta)$ be a $(k \times c)$-matrix with rank c and let Γ_n^* be $(c \times c)$-diagonal matrices with positive diagonal entries and $\Gamma_n^* \to 0$ for $n \to \infty$. Further, let η_n^* be c-dimensional random vectors for which $\Gamma_n^{*^{-1}}(\eta_n^* - \eta)$ is $P_{h(\eta)}$-stochastically bounded, and*

1.) $\Gamma_n^{-1} h'(\eta_n^*)\Gamma_n^* \xrightarrow{P_{h(\eta)}} C(\eta)$,

2.) $U_{n,j}(h(\eta_n^*))\Gamma_n^* h_j''(\eta_n^*)\Gamma_n^* \xrightarrow{P_{h(\eta)}} 0$, $j = 1, \ldots, k$.

Let $\xrightarrow{\mathcal{D}_{h(\eta)}}$ denote the convergence in distribution under the probability law $P_{h(\eta)}$. Then a $\Gamma_n^{*^{-1}}$-consistent Z-estimator $\hat{\eta}_n$ for η exists, satisfying

$$\Gamma_n^{*^{-1}}(\hat{\eta}_n - \eta) \xrightarrow{\mathcal{D}_{h(\eta)}} N_c(0, (B^{*^{-1}}(\eta))^\top \Sigma^*(\eta) B^{*^{-1}}(\eta)) \ ,$$

where $\Sigma^*(\eta) := C^\top(\eta)\Sigma(h(\eta))C(\eta)$ and $B^*(\eta) := C^\top(\eta)B(h(\eta))C(\eta)$.

Furthermore, the log-LR statistic $T_n = 2[\log L(\hat{\vartheta}_n) - \log L(h(\hat{\eta}_n))]$ asymptotically follows a χ^2-law:

$$T_n \xrightarrow{\mathcal{D}_{h(\eta)}} \chi^2_{k-c} \ .$$

Proof: Pruscha [2000, Proposition p. 252, Theorem 4.3 p. 253]. □

Lemma 2.5 *Under the conditions of Theorem 2.3, if the matrices Σ and B are equal, it follows for $\vartheta = h(\eta)$, $\hat{X}_n(\vartheta) := \Gamma_n^{-1}(\hat{\vartheta}_n - \vartheta)$, and $\hat{X}_n^*(\eta) := \Gamma_n^{*^{-1}}(\hat{\eta}_n - \eta)$, that*

1.) $2[\log L(\hat{\vartheta}_n) - \log L(\vartheta)] \ - \ \hat{X}_n^\top(\vartheta)\Sigma(\vartheta)\hat{X}_n(\vartheta) \xrightarrow{P_\vartheta} 0$,

2.) $2[\log L(h(\hat{\eta}_n)) - \log L(h(\eta))] \ - \ \hat{X}_n^{*^\top}(\eta)\Sigma^*(\eta)\hat{X}_n^*(\eta) \xrightarrow{P_{h(\eta)}} 0$.

Thus, for the LR-statistic T_n it holds that

$$T_n \ - \ \hat{X}_n^\top(\vartheta)[\Sigma(\vartheta) - \Sigma(\vartheta)C(\eta)\Sigma^*(\eta)C^\top(\eta)\Sigma(\vartheta)]\hat{X}_n(\vartheta) \xrightarrow{P_{h(\eta)}} 0 \ ,$$

since $\hat{X}_n^*(\eta) \ - \ \Sigma^{*^{-1}}(\eta)C^\top(\eta)\Sigma(h(\eta))\hat{X}_n(h(\eta)) \xrightarrow{P_{h(\eta)}} 0$.

Proof: Pruscha [2000, Proposition (b) p. 195, Corollary p. 249, Theorem 4.3 p.253]. □

The results given above are applied in the following sections to derive the constrained MLEs, the LR statistics and their asymptotic behavior for various models and hypotheses.

3 Two normal samples

3.1 Model and hypotheses

In medical research the comparison of two independent groups is the most popular design. Often, if the outcome is continuous the normality of the data is assumed, or an appropriate data transformation (e.g. logarithm) leads to approximately normal data. The difference between the groups is specified by using the group means. As a rule the group variances are assumed to be equal. This requirement is statistically founded, since in this case the statistical methods are much easier. However, there is an additional reason: The comparison of two means makes sense only for similar variances. Two groups are declared to be equal, if their corresponding means are equal. Certainly, even in the case of equal means of the two groups, these groups are not considered to be equal, if the variances differ substantially.

Let $X_{11}, \ldots, X_{1n_1} \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2)$ and $X_{21}, \ldots, X_{2n_2} \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2)$ be two independent random vectors with equal unknown variances. The mostly used distance measure to discriminate between the two groups is the difference of the means $\theta_d := \mu_1 - \mu_2$. Various authors suggest to use the ratio of the means $\theta_r := \mu_1/\mu_2$ for certain situations (Liu and Weng [1994], Hauschke et al. [1999]). When no information about the data variances is available, the standardized difference $\theta_s := (\mu_1 - \mu_2)/\sigma$ may be used.

In the following it is assumed that the measures θ_d, θ_r and θ_s quantify the inferiority of group 1 compared to group 2.

Thus, for $\theta \in \{\theta_d, \theta_r, \theta_s\}$ the non-inferiority hypotheses are given by

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0 \ ,$$

where θ_0 is a fixed value to be specified in advance (cf. the discussion in Chapter 1).

The group sample means and the pooled standard deviation are denoted by $\overline{x}_1$, $\overline{x}_2$ and s_p, respectively. Furthermore, $(t_{m,\delta})_\alpha$ is the α-quantile of the noncentral t-distribution with m degrees of freedom and noncentrality parameter δ, while $(t_m)_\alpha$ is the α-quantile of the central t-distribution.

3.2 LR test and t-statistics

The classical test for differences in means is the two-sample t-test. The test statistic

$$T_d := \frac{\overline{x}_1 - \overline{x}_2 - \theta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows a noncentral t-distribution with $n_1 + n_2 - 2$ degrees of freedom and noncentrality parameter

$$\delta_d := \frac{\mu_1 - \mu_2 - \theta_0}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\theta_d - \theta_0}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \ . \tag{3.1}$$

At the boundary of the null space ($\theta_d = \theta_0$) the statistic T_d follows a t-distribution. The null hypothesis H_0 is rejected at level α for $T_d < (t_{n_1+n_2-2})_\alpha$. This test is the uniformly most powerful unbiased test (Lehmann [1986]), and it is equivalent to the LR test, since the LR-statistic and T_d are equivalent. This is shown in the following Lemma.

Lemma 3.1 *If $\overline{x}_1 - \overline{x}_2 < \theta_0$ holds, the LR-statistic for θ_d is a strictly monotone transformation of the t-statistic T_d.*

Proof: The unconstrained MLE for μ_1, μ_2 and σ^2, respectively, is given by $\overline{x}_1$, $\overline{x}_2$, and

$$\hat{\sigma}^2 := \frac{n_1 + n_2 - 2}{n_1 + n_2}\ s_p^2 \ .$$

Due to Theorem 2.1 the MLEs constrained to H_0 are located at the boundary of the null space and are given (using the results of Mood et al. [1974, Ch. IX, 4.3]) by

$$\begin{aligned}
\mu_1^* &= \frac{n_1\overline{x}_1 + n_2(\overline{x}_2 + \theta_0)}{n_1 + n_2} \ , \\
\mu_2^* &= \frac{n_1(\overline{x}_1 - \theta_0) + n_2\overline{x}_2}{n_1 + n_2} \ , \\
\sigma^{2^*} &= \hat{\sigma}^2 + \frac{n_1 n_2}{(n_1 + n_2)^2}(\overline{x}_1 - \overline{x}_2 - \theta_0)^2 \ .
\end{aligned}$$

Thus, for $\overline{x}_1 - \overline{x}_2 < \theta_0$ the LR-statistic λ is given by

$$\lambda = \left[1 + \frac{n_1 n_2}{n_1 + n_2}\frac{(\overline{x}_1 - \overline{x}_2 - \theta_0)^2}{(n_1 + n_2 - 2)s_p^2}\right]^{-\frac{n_1+n_2}{2}} = \left[1 + \frac{T_d^2}{n_1 + n_2 - 2}\right]^{-\frac{n_1+n_2}{2}} \ .$$

□

Using the ratio θ_r as the distance measure, Sasabuchi [1980] has shown that the LR test is equivalent to a t-test as well. The test statistic is

$$T_r := \frac{\overline{x}_1 - \theta_0 \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{\theta_0^2}{n_2}}} \sim t_{n_1+n_2-2,\delta_r} ,$$

where

$$\delta_r := \frac{\mu_1 - \theta_0 \mu_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{\theta_0^2}{n_2}}} = \frac{\theta_r - \theta_0}{\frac{\sigma}{\mu_2} \sqrt{\frac{1}{n_1} + \frac{\theta_0^2}{n_2}}} .$$

For $\theta_r = \theta_0$ the distribution simplifies to a central t-distribution. Thus, the null hypothesis is rejected for $T_r < (t_{n_1+n_2-2})_\alpha$.

For the standardized difference θ_s the test statistic T_d is used with $\theta_0 = 0$:

$$T_s := \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} .$$

This statistic is noncentrally t-distributed with $n_1 + n_2 - 2$ degrees of freedom and noncentrality parameter

$$\delta_s := \frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\theta_s}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} . \tag{3.2}$$

In order to make a test decision the α-quantile of the noncentral t-distribution has to be calculated. The null hypothesis is rejected for $T_s < (t_{n_1+n_2-2,\delta(\theta_0)})_\alpha$ where $\delta(\theta_0)$ is the noncentrality parameter from (3.2) with $\theta_s = \theta_0$.

Lehmann [1986, p. 294] has shown that this noncentral t-test is the uniformly most powerful invariant test with respect to the group of scale transformations. Note, that this test is different from the LR test. Since the difference in means is constrained in terms of the standard deviation, numerical calculations are required to determine the constrained MLE for the LR test. However, the LR test for the standardized difference is omitted here, since any reasonable statistical tests has to be invariant with respect to the choice of the measuring scale.

3.3 Power and sample size calculation

It is possible to calculate the power for a given sample size and to calculate the minimum sample size for a given power for all three distance measures, since the distribution of the test statistics T_d, T_r and T_s is known for normal data. This will be briefly indicated in the following.

The test statistic T_d is noncentrally t-distributed with $n_1 + n_2 - 2$ degrees of freedom and the noncentrality parameter δ_d given in (3.1). Thus, the power for specified sample sizes n_1 and n_2 and a distance θ_d ($< \theta_0$) is calculated as

$$1 - \beta := P_{\theta_d}(T_d < (t_{n_1+n_2-2})_\alpha) = F_{n_1+n_2-2,\delta_d}((t_{n_1+n_2-2})_\alpha) \ , \tag{3.3}$$

where $F_{m,\delta}(x)$ is the cumulative distribution function of the noncentral t-distribution with m degrees of freedom and noncentrality parameter δ which is available in many statistical software packages.

In planning a clinical trial the required sample size has to be calculated to obtain a given power $1 - \beta$. If the ratio $\epsilon := n_1/n_2$ is fixed, the power in (3.3) is isotonic in n_1. Thus, the minimal sample size n_1^* is given by

$$n_1^* = \min\{n_1 \in \mathbb{N} : \ F_{n_1+n_2-2,\delta_d}((t_{n_1+n_2-2})_\alpha) \geq 1 - \beta\} \ , \tag{3.4}$$

where n_2 is replaced by n_1/ϵ. If no statistical software package is available for calculating the noncentral t-distribution, the following approximation of the α-quantile of the noncentral t-distribution can be used (Johnson and Welch [1940, p. 207]). For large sample sizes, as $\min\{n_1, n_2\} \to \infty$,

$$\begin{aligned} (t_{n_1+n_2-2})_\alpha &= u_\alpha + o(1) \ , \\ (t_{n_1+n_2-2,\delta})_\alpha &= \delta + u_\alpha \sqrt{1 + \frac{1}{2(n_1+n_2-2)}(\delta^2 - u_\alpha^2)} + o(1) \ , \end{aligned} \tag{3.5}$$

where u_α denotes the α-quantile of the standard normal distribution. Therefore, an approximation for n_1^* from (3.4) is obtained by means of the normal distribution. With $\Delta_d := \frac{\theta_d - \theta_0}{\sigma}$ (and thus $\delta_d = \Delta_d \sqrt{\frac{n_1}{1+\epsilon}}$) the requirement $(t_{n_1+n_2-2,\delta_d})_{1-\beta} \geq (t_{n_1+n_2-2})_\alpha$ is asymptotically equivalent to

$$\Delta_d \sqrt{\frac{n_1}{1+\epsilon}} + u_{1-\beta} \sqrt{1 + \frac{1}{2(n_1(1+\epsilon^{-1})-2)}(\Delta_d^2 \frac{n_1}{1+\epsilon} - u_\alpha^2)} \geq u_\alpha \ ,$$

which implies that

$$n_1 \geq (1+\epsilon) \frac{\left(u_\alpha - u_{1-\beta} \sqrt{1 + \frac{\Delta_d^2}{2(1+\epsilon^{-1})(1+\epsilon)}} \right)^2}{\Delta_d^2} \ + \ o(1) \ . \tag{3.6}$$

Hence, by (3.6) the total sample size $N := n_1 + n_2 = n_1(1+\epsilon^{-1})$ depends on ϵ only via the term $(1+\epsilon)(1+\epsilon^{-1})$. Thus, the optimal (in terms of a minimal N) group allocation is given for $\epsilon = 1$, since the term $(1+\epsilon)(1+\epsilon^{-1})$ is minimal for $\epsilon = 1$ and the function $f(x) = x\left(-a - b\sqrt{1+\frac{c}{2x}}\right)^2$ is isotonic for positive constants a, b, c and $x > 0$.

Analogously, for θ_r $(< \theta_0)$ and fixed sample sizes n_1 and n_2 the power of the t-test is given by

$$1 - \beta := P_{\theta_r}(T_r < (t_{n_1+n_2-2})_\alpha) = F_{n_1+n_2-2,\delta_r}((t_{n_1+n_2-2})_\alpha) \ .$$

For a given power $1-\beta$ and an allocation ϵ, the minimal n_1^* is also obtained using (3.4). However, the noncentrality parameter δ_d in (3.4) has to be replaced by δ_r. An approximation for n_1^* is given by (analogously to (3.6))

$$n_1 \geq (1+\epsilon\theta_0^2)\frac{\left(u_\alpha - u_{1-\beta}\sqrt{1+\frac{\Delta_r^2}{2(1+\epsilon^{-1})(1+\epsilon\theta_0^2)}}\right)^2}{\Delta_r^2} \ ,$$

where $\Delta_r := \mu_2\frac{\theta_r-\theta_0}{\sigma}$ (and thus $\delta_r = \Delta_r\sqrt{\frac{n_1}{1+\epsilon\theta_0^2}}$).

In contrast to the difference θ_d, the 1:1 allocation $n_1 = n_2$ is not optimal when using the ratio as the distance measure. To minimize the overall sample size for a fixed power $1-\beta$, the allocation $\epsilon = \theta_0^{-1}$ has to be chosen, since this is the minimum of $(1+\epsilon^{-1})(1+\epsilon\theta_0^2)$.

Using the standardized difference θ_s as the distance measure, the power for θ_s $(< \theta_0)$ is calculated by

$$1 - \beta := P_{\theta_s}(T_r < (t_{n_1+n_2-2,\delta(\theta_0)})_\alpha) = F_{n_1+n_2-2,\delta_s}((t_{n_1+n_2-2,\delta(\theta_0)})_\alpha) \ ,$$

where $\delta(\theta_0)$ is the noncentrality parameter from (3.2) with $\theta_s = \theta_0$. With (3.5) an approximation for the required sample size n_1^* is given by

$$\min\{n_1 \in \mathbb{N} : \ n_1 \geq (1+\epsilon)\frac{(u_\alpha - u_{1-\beta})^2}{(\theta_s - \theta_0)^2}\} \ .$$

As for the difference, asymptotically the samples have to be equally sized $(\epsilon = 1)$ to maximize the power for θ_s.

4 Three normal samples

Several clinical trials aim at comparing three or more parallel groups. Some reasons are mentioned in Chapter 1. For one-sided hypotheses and more than two groups the statistical theory is based on methods of order restricted statistical inference which was extensively developed since the early 1950s. Barlow et al. [1972] have summarized much of the early work. For k independent normally distributed groups with means $\mu = (\mu_1, \ldots, \mu_k)$ they considered the null hypothesis $H_0 : \mu_1 = \ldots = \mu_k$ versus various types of ordered alternatives. Robertson et al. [1988] extended this work. Additionally, hypotheses of the type

$$H_0 : \mu \text{ is isotonic with respect to } \preceq \quad \text{vs.} \quad \neg H_0 \ ,$$

are considered, where $\preceq$ is a partial ordering of μ. Robertson et al. developed the LR test for different partial orderings.

Another possibility to deal with more than two groups is the application of multiple comparison procedures (pairwise comparisons) which are more commonly used in clinical research. The following sections will show that the LR test for some of the hypotheses introduced in Chapter 1 is indeed equivalent to a multiple comparison procedure. For other hypotheses the LR principle cannot be reduced to pairwise comparisons and, thus, yields different statistical tests. This will be investigated in detail in the subsequent sections.

4.1 Model and hypotheses

Throughout the following, a sample of independent normally distributed random variables of three homoscedastic groups is given by

$$y_{ij} \sim N(\mu_i, \sigma^2) \quad (i = 1, 2, 3 \ ; \ j = 1, \ldots, n_i) \ ,$$

where $\mu = (\mu_1, \mu_2, \mu_3)$ is the vector of the group means, σ the common standard deviation, and n_i the sample size of group i, $i = 1, 2, 3$.

Following the hypotheses a) - c) stated in Chapter 1, three kinds of null hypotheses are

investigated specified by the differences of means:

$$\begin{aligned} H_0^a &: \mu_1 - \mu_2 \geq \theta_1 \ \vee \ \mu_1 - \mu_3 \geq \theta_2 \ , \\ H_0^b &: \mu_1 - \mu_2 \geq \theta_1 \ \wedge \ \mu_1 - \mu_3 \geq \theta_2 \ , \\ H_0^c &: \mu_1 - \mu_2 \geq \theta_1 \ \vee \ \mu_2 - \mu_3 \geq \theta_2 \ . \end{aligned} \tag{4.1}$$

Without changing the likelihood and hence the testing problem, the two equivalence margins θ_1, θ_2 can be set to zero when adding θ_1 to the values of group 2 and θ_2 to the values of group 3. Hence, throughout the following we assume $\theta_1 = \theta_2 = 0$.

Of course, when other distance measures are specified, shifting of the margins may no longer be possible (e.g. for the standardized difference). Nevertheless, in this chapter we restrict to the difference, since it is the most commonly used distance measure in clinical research.

4.2 Multiple comparison procedures

The testing problems (4.1) can be solved by using two pairwise two-sample tests. When applying multiple comparison procedures, the common type I error probability has to be taken into consideration, i.e. the probability under H_0 to reject at least one of the tests. It depends heavily on the arrangement of the hypotheses whether a level adjustment is needed or not.

Referring to the hypotheses a) and c), the null hypotheses are rejected if both pairwise comparisons are rejected. This testing procedure is called an *intersection-union test* (IUT), since the null hypothesis is the union of two subhypotheses and the alternative hypothesis is the intersection of the complements of the subhypotheses. Berger [1982,Theorem 1] showed that intersection-union tests keep the nominal size α with each of the multiple tests carried out as a level α test. Therefore, the global hypotheses H_0^a and H_0^c are rejected if the pairwise two-sample t-tests are both rejected without level adjustment. It will be seen in the next section that this procedure is nearly equivalent to the LR test.

Focussing on the hypothesis H_0^b, the testing problem is different. One way to deal with this situation is to reject H_0^b if at least one of the two pairwise tests is rejected. Here, a level adjustment is necessary in order to guarantee the global level. Various authors suggested several strategies to handle this problem. This includes so called *single step* procedures, i.e. all pairwise comparisons are carried out simultaneously, as well as *step up/down* procedures. For these procedures the pairwise comparisons are arranged (in order of their p-values) and the subhypotheses are rejected, iteratively, until given criteria are fulfilled. For the general theory confer Dunnett and Tamhane [1997], Hsu [1996],

D'Agostino and Heeren [1991], or Hochberg and Tamhane [1987]. We give a short overview over the most commonly used procedures in the particular situation of H_0^b.

Throughout the following, let p_1, p_2 be the p-values of the pairwise t-tests for the null hypotheses $H_{0_1}^b : \mu_1 - \mu_2 \geq \theta_1$ and $H_{0_2}^b : \mu_1 - \mu_3 \geq \theta_2$, and $p_{(1)}, p_{(2)}$ the smaller and the larger of both p-values, respectively.

The easiest single step method to adjust the global level α for the hypothesis H_0^b is the *Bonferroni adjustment*. Here α is evenly divided to each pairwise comparison, i.e. H_0^b is rejected if $p_{(1)} < \frac{\alpha}{2}$. Holm [1979] suggested a step down procedure improving Bonferroni's adjustment. Note that for H_0^b both procedures are equal.

A further improvement was introduced by Hochberg [1988] for multiple tests with independent test statistics. This step up procedure applied to the hypothesis H_0^b starts with the larger p-value. If $p_{(2)} < \alpha$, both subhypotheses $H_{0_1}^b, H_{0_2}^b$ are rejected. If $p_{(2)}$ is larger than α and $p_{(1)} < \frac{\alpha}{2}$, then only the subhypothesis corresponding to the smaller p-value is rejected. Otherwise no subhypothesis is rejected. Applied to H_0^b the hypothesis is rejected, if $p_{(2)} < \alpha$ or $p_{(1)} < \frac{\alpha}{2}$. Note that the test statistics are not independent for H_0^b. Hence, Hochberg's procedure applied to our setting does not guarantee to keep the nominal level α. Nevertheless, we will consider this method in the following, since we found numerically quite satisfactory results.

Dunnett [1955] has introduced a single step approach which provides simultaneous one-sided confidence intervals for normally distributed data. The multivariate t-distribution is used with a particular correlation matrix (for explicit formulae confer Dunnett's work). Applying Dunnett's procedure (e.g. using the SAS-function *PROBMC*), the hypothesis H_0^b is rejected if the upper confidence limit for $\mu_1 - \mu_2$ is smaller than θ_1 or the upper confidence limit for $\mu_1 - \mu_3$ is smaller than θ_2.

It will be seen that the LR test for H_0^b is different from the above mentioned pairwise procedures, even if the differences are small in practice.

4.3 Likelihood ratio statistic

The general methodology deriving the LR test is embraced by the term *order restricted inference* (see e.g. Robertson et al. [1988]). In this section the required formulae are given for the hypotheses H_0^a, H_0^b and H_0^c.

The likelihood function is given by

$$L(\mu, \sigma) = (2\pi\sigma^2)^{\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2\right] ,$$

where $N := n_1 + n_2 + n_3$.

Robertson et al. [1988, p. 63] showed for $H \subset \mathbb{R}^3$ that

$$\arg\max_{\mu\in H} L(\mu,\sigma) = \arg\max_{\mu\in H} L(\mu,1) =: \mu_H^* = (\mu_{H,1}^*, \mu_{H,2}^*, \mu_{H,3}^*) \ ,$$

$$\arg\max_{\sigma} L(\mu,\sigma) = \arg\max_{\sigma} L(\mu_H^*,\sigma) = \left[\frac{\sum_{i=1}^3\sum_{j=1}^{n_i}(y_{ij}-\mu_{H,i}^*)^2}{n_1+n_2+n_3}\right]^{\frac{1}{2}} =: \sigma_H^* \ .$$

It follows that the LR for H_0 vs. H_1 is given by

$$\lambda = \left[\frac{\sigma^*_{H_0\cup H_1}}{\sigma^*_{H_0}}\right]^N \ \Leftrightarrow \ \lambda^{\frac{2}{N}} = \frac{SS_W}{\sum_{i=1}^3\sum_{j=1}^{n_i}(y_{ij}-\mu_i^*)^2} \ , \tag{4.2}$$

where $SS_W := \sum_{i=1}^3\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$, $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$, and μ_i^* is the i-th component of the vector $\arg\max_{\mu\in H_0} L(\mu,1)$.

The denominator from the right hand side of (4.2) is equal to

$$SS_W + \sum_{i=1}^3\sum_{j=1}^{n_i}(\overline{y}_i-\mu_i^*)^2 - 2\sum_{i=1}^3\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)(\mu_i^*-\overline{y}_i) \ .$$

The last term is zero (Robertson et al. [1988, Th. 1.3.6]), therefore

$$\lambda^{\frac{2}{N}} = \frac{SS_W}{SS_W + Z} \ ,$$

with $Z := \sum_{i=1}^3 n_i(\overline{y}_i-\mu_i^*)^2$. Since $(1-x)/x$ is a monotone transformation of $(0,1)$, the test statistic

$$S = \frac{Z}{SS_W} \tag{4.3}$$

can be used instead of $\lambda^{\frac{2}{N}}$. The determination of the constrained MLE μ_i^* as $\arg\max_{\mu\in H_0} L(\mu,1)$ is equivalent to determine $\arg\min_{\mu\in H_0} Z$.

Hypothesis H_0^a

Here we derive the test statistic and its asymptotic null distribution for

$$H_0^a : \mu_1-\mu_2 \geq 0 \ \vee \ \mu_1-\mu_3 \geq 0 \quad \text{vs.} \quad \neg H_0^a \ .$$

If $\overline{y}_1 \geq \min\{\overline{y}_2,\overline{y}_3\}$, it follows that $Z = S = 0$. If $\overline{y}_1 < \min\{\overline{y}_2,\overline{y}_3\}$, the statistic Z is calculated by

$$Z = \begin{cases} \frac{n_1n_2}{n_1+n_2}(\overline{y}_1-\overline{y}_2)^2 & \text{if } \sqrt{\frac{n_2}{n_1+n_2}}(\overline{y}_2-\overline{y}_1) < \sqrt{\frac{n_3}{n_1+n_3}}(\overline{y}_3-\overline{y}_1) \\ \frac{n_1n_3}{n_1+n_3}(\overline{y}_1-\overline{y}_3)^2 & \text{else} \end{cases} \ ,$$

since

$$\min_{\{\mu_1\geq\mu_2\}\cup\{\mu_1\geq\mu_3\}} Z = \min\{\min_{\mu_1\geq\mu_2} Z, \min_{\mu_1\geq\mu_3} Z\} \ ,$$

and $\arg\min_{\mu_1\geq\mu_g} Z$ is given by

$$\mu_1^* = \mu_g^* = \frac{n_1\overline{y}_1 + n_g\overline{y}_g}{n_1+n_g}$$

for $g = 2, 3$ (cf. Robertson et al. [1988, p. 63]).

The LR test rejects H_0^a for small values of S. The probability that the test statistic (4.3) exceeds a margin $c > 0$ is given by

$$P(Z > c\ SS_W) = P\left(\sqrt{cSS_W} < \min\left\{\sqrt{\tfrac{n_1n_2}{n_1+n_2}}(\overline{y}_2-\overline{y}_1), \sqrt{\tfrac{n_1n_3}{n_1+n_3}}(\overline{y}_3-\overline{y}_1)\right\}\right).$$

Since the distribution of SS_W is independent of the means μ_i, the probability $P(Z > c\ SS_W)$ is isotonic in μ_2 and μ_3 and antitonic in μ_1. The worst case, i.e. the maximal probability under H_0^a, results if two of the three means are equal and the third mean is infinity (cf. Berger [1982, Th. 2]). Thus it is

$$\begin{aligned}
\max_{\mu\in H_0^a} P_\mu(Z > c\ SS_W) &= \lim_{\mu_3\to\infty} P_{\mu_1=\mu_2}(Z > c\ SS_W) \\
&= \lim_{\mu_2\to\infty} P_{\mu_1=\mu_3}(Z > c\ SS_W) \\
&= P_{\mu_1=\mu_2}(\sqrt{c\ SS_W} < \sqrt{\tfrac{n_1n_2}{n_1+n_2}}(\overline{y}_2-\overline{y}_1)) \\
&= P_{\mu_1=\mu_2}(\sqrt{c(N-3)} < \sqrt{\tfrac{n_1n_2(N-3)}{n_1+n_2}}\ \frac{\overline{y}_2-\overline{y}_1}{\sqrt{SS_W}}).
\end{aligned}$$

The random variable on the right hand side of the last term is t_{N-3}-distributed. Therefore, the hypothesis H_0^a is rejected for $(N-3)Z > SS_W\ (t_{N-3})_{1-\alpha}$.

Hypothesis H_0^c

Following the same arguments as for H_0^a, the test statistic (4.3) for

$$H_0^c : \mu_1 - \mu_2 \geq 0\ \vee\ \mu_2 - \mu_3 \geq 0 \quad \text{vs.} \quad \neg H_0^c$$

is zero if $\overline{y}_1 \geq \overline{y}_2$ or $\overline{y}_2 \geq \overline{y}_3$. If $\overline{y}_1 < \overline{y}_2 < \overline{y}_3$, the test statistic Z is calculated as

$$Z = \begin{cases} \frac{n_1n_2}{n_1+n_2}\ (\overline{y}_1-\overline{y}_2)^2 & \text{if } \sqrt{\frac{n_1}{n_1+n_2}}\ (\overline{y}_2-\overline{y}_1) < \sqrt{\frac{n_3}{n_2+n_3}}\ (\overline{y}_3-\overline{y}_2) \\ \frac{n_2n_3}{n_2+n_3}\ (\overline{y}_2-\overline{y}_3)^2 & \text{else} \end{cases} .$$

Further, the worst case under H_0^c is given by

$$\max_{\mu\in H_0^c} P_\mu(Z > c\ SS_W) = P_{\mu_1=\mu_2}(\sqrt{c(N-3)} < \sqrt{\tfrac{n_1n_2N}{n_1+n_2}}\ \frac{\overline{y}_2-\overline{y}_1}{\sqrt{SS_W}}) \ .$$

Therefore, the hypothesis H_0^c is rejected for $(N-3)Z > SS_W\ (t_{N-3})_{1-\alpha}$.

Remark 4.1 *The LR test is equivalent to the IUT if for the IUT test the two-sample variance estimates are replaced by the pooled three-sample variance estimates. The global hypotheses H_0^a and H_0^c are rejected if both the subhypotheses are rejected at level α by using the pairwise two-sample t-tests. This leads to an improvement over the pairwise estimation using the two-sample pooled standard deviation due to the larger number of $N-3$ degrees of freedom.*

Hypothesis H_0^b

This hypothesis is a particular case of a simple tree hypothesis for $k > 2$ groups (Robertson et al. [1988]). In the following the formulae for the particular case of three homoscedastic groups will be derived.

The test statistic for

$$H_0^b : \mu_1 - \mu_2 \geq 0 \ \wedge \ \mu_1 - \mu_3 \geq 0 \quad \text{vs.} \quad \neg H_0^b$$

is (cf. Robertson et a. [1988, p. 65 (2.2.13)] given by

$$S = 1 - \lambda^{\frac{2}{N}} \ . \tag{4.4}$$

The constrained MLEs μ_i^* $(i = 1, 2, 3)$ are $\mu_i^* = \overline{y}_i$, if $\overline{y}_1 \geq \max\{\overline{y}_2, \overline{y}_3\}$. If $\overline{y}_1 < \overline{y}_2$ and $\overline{y}_2 > \overline{y}_3$ (for the case $\overline{y}_2 \leq \overline{y}_3$ exchange $\overline{y}_2$ and $\overline{y}_3$), the constrained MLEs are calculated as follows (cf. Robertson et al. [1988, p. 19]). If $\overline{y}_{12} \geq \overline{y}_3$, we obtain

$$\mu_1^* = \mu_2^* = \overline{y}_{12} := \frac{n_1\overline{y}_1 + n_2\overline{y}_2}{n_1 + n_2} \ , \quad \mu_3^* = \overline{y}_3 \ ,$$

and if $\overline{y}_{12} < \overline{y}_3$,

$$\mu_1^* = \mu_2^* = \mu_3^* = \overline{y}_{123} := \frac{n_1\overline{y}_1 + n_2\overline{y}_2 + n_3\overline{y}_3}{n_1 + n_2 + n_3} \ ,$$

where $\overline{y}_{12}$ and $\overline{y}_{123}$ are the weighted arithmetic means of group 1 and 2, and 1, 2 and 3, respectively.

Since $P_{H_0^b}(S > c) \leq P_{\mu_1=\mu_2=\mu_3}(S > c)$ for all $c > 0$ (cf. Robertson et al. [1988, p. 68, 69]), the critical value c to reject the null hypothesis is determined by the equation

$$\begin{aligned} P_{H_0^b}(S > c) &= p_{[1]} \ P(B_{1,(n_1+n_2+n_3-3)/2} > c) \\ &\quad + p_{[2]} \ P(B_{1/2,(n_1+n_2+n_3-3)/2} > c) \\ &= \alpha \ . \end{aligned}$$

where $B_{\cdot,\cdot}$ is the Beta-distribution. The values $p_{[i]}$ are the probabilities for the following event: The vector $\mu^* = (\mu_1^*, \mu_2^*, \mu_3^*)$ includes exactly i different values.

The probabilities $p_{[i]}$ can be calculated using Formula (10) from Childs [1967]: For two normally distributed random variables Z_1 and Z_2 with correlation coefficient ρ and $E(Z_1) = E(Z_2) = 0$, it is

$$P(Z_1 < 0, Z_2 < 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin \rho \ .$$

Since $p_{[1]} + p_{[2]} + p_{[3]} = 1$ and

$$\begin{aligned} p_{[1]} &= P(\overline{y}_3 - \overline{y}_2 < 0 \ , \ \frac{n_1\overline{y}_1 + n_2\overline{y}_2}{n_1 + n_2} - \overline{y}_3 < 0) \\ &\quad + P(\overline{y}_2 - \overline{y}_3 < 0 \ , \ \frac{n_1\overline{y}_1 + n_3\overline{y}_3}{n_1 + n_3} - \overline{y}_2 < 0) \ , \\ p_{[3]} &= P(\overline{y}_3 - \overline{y}_1 < 0 \ , \ \overline{y}_2 - \overline{y}_1 < 0) \ , \end{aligned}$$

$p_{[1]}$ and $p_{[2]}$ are calculated by

$$\begin{aligned} p_{[1]} &= \frac{1}{2} + \frac{1}{2\pi} \left(\arcsin \left[-\sqrt{\frac{n_1 + n_2 + n_3}{(1 + \frac{n_3}{n_2})(n_1 + n_2)}} \right] \right. \\ &\quad \left. + \arcsin \left[-\sqrt{\frac{n_1 + n_2 + n_3}{(1 + \frac{n_2}{n_3})(n_1 + n_3)}} \right] \right) \ , \\ p_{[2]} &= 1 - \frac{1}{4} - \frac{1}{2\pi} \arcsin \left[\left((1 + \frac{n_1}{n_2})(1 + \frac{n_1}{n_3}) \right)^{-1/2} \right] - p_{[1]} \ . \end{aligned}$$

To sum up, the LR test for the hypotheses H_0^a and H_0^c is reduced to commonly known two-sample procedures. In case of hypothesis H_0^b, Robertson et al. [1988] developed the LR test for more than two groups which results in the above mentioned formulae for three groups.

4.4 Power investigation

It is shown that for H_0^a and H_0^c the LR principle leads to the IUT, i.e. the pairwise two-sample t-tests at level α have to be calculated, using the data of all the three groups to estimate the pooled standard deviation. A power investigation is omitted for the hypotheses H_0^a and H_0^c, since the LR test is equivalent to standard methods extensively investigated in the past. Therefore, it is focussed on hypothesis H_0^b. The interesting question is whether the power is increased by the LR test in comparison to the pairwise comparison procedures described above.

The power of the LR test and its pairwise competitors are computed by simulations (100,000 replications in each scenario) for different sample sizes and means μ_1, μ_2, μ_3.

Table 4.1: The simulated power (times 100) of the LR test and its corresponding pairwise test procedures using Bonferroni's, Dunnett's and Hochberg's level adjustment for different parameter constellations.

n_1	n_2	n_3	μ_1	μ_2	μ_3	LR test	Bonferroni	Dunnett	Hochberg
20	20	20	0	0	0.97	85.6	85.3	86.4	85.3
30	30	30	0	0	0.78	85.2	85.0	86.0	85.0
40	40	40	0	0	0.67	84.8	84.5	85.6	84.5
50	50	50	0	0	0.59	84.1	83.9	85.0	83.9
60	60	60	0	0	0.55	84.9	84.7	85.7	84.7
80	80	80	0	0	0.48	85.5	85.3	86.3	85.3
100	100	100	0	0	0.42	84.9	84.7	85.7	84.7
40	20	20	0	0	0.84	85.6	86.0	86.7	86.1
60	30	30	0	0	0.69	85.7	86.2	86.7	86.2
80	40	40	0	0	0.59	85.7	86.2	86.8	86.2
100	50	50	0	0	0.53	85.9	86.4	86.9	86.4
20	40	40	0	0	0.83	84.6	84.8	85.6	84.9
30	60	60	0	0	0.67	84.4	84.7	85.4	84.7
40	80	80	0	0	0.58	84.3	84.6	85.3	84.6
50	100	100	0	0	0.52	85.2	85.4	86.1	85.4
20	20	20	-0.78	0	0	85.3	82.7	84.0	84.2
30	30	30	-0.63	0	0	84.4	81.7	83.0	83.2
40	40	40	-0.55	0	0	85.0	82.4	83.7	83.9
50	50	50	-0.48	0	0	84.6	82.0	83.3	83.4
60	60	60	-0.44	0	0	84.0	81.4	82.7	82.9
80	80	80	-0.38	0	0	84.7	82.2	83.4	83.6
100	100	100	-0.34	0	0	85.1	82.6	83.8	84.0
40	20	20	-0.64	0	0	85.1	81.7	82.5	83.2
60	30	30	-0.52	0	0	84.3	81.0	81.8	82.4
80	40	40	-0.45	0	0	85.4	82.4	83.1	83.7
100	50	50	-0.41	0	0	85.5	82.5	83.2	83.8
20	40	40	-0.58	0	0	84.7	81.9	82.9	83.3
30	60	60	-0.47	0	0	84.3	81.6	82.6	82.9
40	80	80	-0.41	0	0	84.6	81.9	82.9	83.3
50	100	100	-0.37	0	0	85.1	82.6	83.5	83.9

All pairwise comparisons are calculated applying Bonferroni's, Dunnett's and Hochberg's procedure. The data of all three groups are included to estimate the pooled standard deviation for the pairwise comparisons, simultaneously.

Table 4.1 shows the power for sample sizes of 20 - 100 per group; balanced and unbalanced cases are incorporated. According to the medical problem 2 (one treatment T compared with two controls C_1, C_2) and 3 (two treatments T_1, T_2 compared with one control C) of Chapter 1, two different constellations are investigated. First, suppose that T is better than C_1 but not better than C_2, or suppose that T_1 is better than C but T_2 is not better than C. Then in both cases the power for $\mu_1 = \mu_2 < \mu_3$ may be of interest. W.l.o.g. $\mu_1 = \mu_2 = 0$ and $\mu_3 > 0$. Another constellation is given by $\mu_1 < \mu_2 = \mu_3$, supposing T is better than two equal controls, or supposing T_1 is equal to T_2 and both are better than C. W.l.o.g. $\mu_2 = \mu_3 = 0$ and $\mu_1 < 0$. We have chosen those values for μ_3 and μ_1, respectively, where the power is approximately 0.85.

From Table 4.1 we draw that in general the power differences between the LR test and the pairwise procedures are small. For $\mu_1 = \mu_2 < \mu_3$ the LR test, Bonferroni's and Hochberg's procedure are almost indistinguishable, whereas Dunnett's procedure gives a slight improvement in power. The differences in power are slightly larger for the case $\mu_1 < \mu_2 = \mu_3$. Here the LR test leads to an improvement.

As an overall conclusion the LR test is comparable to the best of the pairwise procedures with respect to power. For the case $\mu_1 < \mu_2 = \mu_3$ the power can be slightly increased using the LR test.

As mentioned above applying Hochberg' procedure it is not guaranteed that the nominal level α is strictly kept. Nevertheless, the comparison to the other procedures is valid, since we found for the investigated parameter constellations (with $\mu_1 = \mu_2 = \mu_3$) that the level of Hochberg's procedure ranges between 0.045 and 0.049.

In a more extensive simulation study we found that the differences for smaller power values (e.g. 0.3 or 0.5) are negligible (not displayed).

In this chapter we assumed homoscedasticity. It would be a goal in further investigations to derive and analyze the LR test in case of unequal group variances.

5 Two binomial samples

5.1 Introduction and hypotheses

The most common set-up in non-inferiority trials is the comparison of two treatments, where the primary endpoint is a dichotomous quantity, such as a success or failure rate. Several statistical methods have been suggested during the last years. See Chan [1998], Farrington and Manning [1990] or Roebruck and Kühn [1995] for a survey on testing methods for the difference of the failure rates. However, there is a controversial discussion on how to measure non-inferiority properly. In addition to the difference $\theta_{DI} := \vartheta_1 - \vartheta_2$, various authors suggest the relative risk $\theta_{RR} := \vartheta_1/\vartheta_2$ or the odds ratio $\theta_{OR} := \vartheta_1 (1 - \vartheta_2) / (\vartheta_2 (1 - \vartheta_1))$. The ASSENT-2 trial [1999] compared two thrombolytic therapies with respect to 30-days mortality rates. Here θ_{DI} as well as θ_{RR} were evaluated.

Proper null hypotheses associated with these quantities are of the form $H_0 \colon \theta_{DI,RR,OR} \geq \theta_0$, where θ_0 is a positive quantity to be specified. Typical values are $\theta_0 = 0.1, 0.15, 0.2$ for the difference and $\theta_0 = 1.1, 1.2, 1.5$ for the relative risk or odds ratio, say (Committee for Proprietary Medicinal Products [1997, 1999], FDA [1992, 1998], ILAE [1998], InTIME-II [2000], Moliterno and Topol [2000]). Phillips [2003] considered hypotheses with linear inequalities $\vartheta_1 \geq a + b\ \vartheta_2$ for fixed a and b and provided an asymptotic test (based on a standardized z-statistic with unpooled variance estimates).

Recently, Röhmel and Mansmann [1999b] argued forcefully that even more general hypotheses are of interest. These authors considered various types of hypotheses which can be described as

$$H_0 \colon \vartheta_1 \geq h(\vartheta_2) \text{ versus } H_1 \colon \vartheta_1 < h(\vartheta_2). \tag{5.1}$$

Here h is an increasing curve $h \colon [0,1] \to [0,1]$ which has to be specified in advance. This includes in particular the above mentioned quantities for

$$h_{DI}(\vartheta_2) = \vartheta_2 + \theta_0, \quad h_{RR}(\vartheta_2) = \vartheta_2\, \theta_0, \quad h_{OR}(\vartheta_2) = \frac{\theta_0}{\theta_0 + \vartheta_2^{-1} - 1}\ , \tag{5.2}$$

or Phillips' [2003] hypotheses.

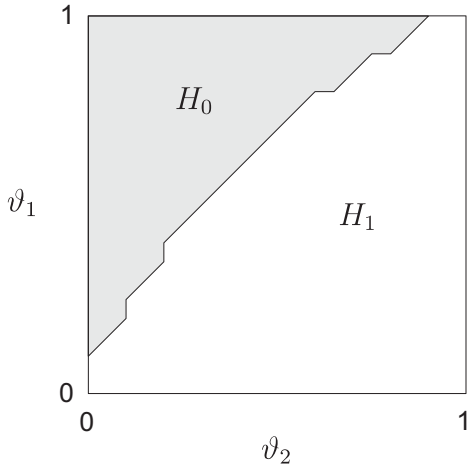

Figure 5.1: Parameter space FDA.

More generally, h might take into account that different measures of discrepancy as well as different values of θ_0 have to be combined in one quantity, depending on the underlying response rate.

Based on recent guidelines of the FDA [1992] and CPMP [1997, 1999], Röhmel and Mansmann [1999b] (see also Bristol [1996]) considered such curves h. Some of them may even be discontinuous, while always being increasing.

As an example, the FDA [1992] requires that non-inferiority can be claimed if the two-sided 95% confidence interval around the difference in response rates must be within θ_0, with

$$\theta_0 = \begin{cases} 20\% & & < 80\% \\ 15\% & \text{if } \max\left\{\hat{\vartheta}_1, \hat{\vartheta}_2\right\} & \in [80\%, 90\%) \\ 10\% & & \geq 90\% \end{cases} .$$

If this rule is applied by replacing the observed rates with the true rates and extrapolating the margins in a symmetric way for small rates (< 0.5), it results in the curve displayed in Figure 5.1. For a careful discussion and other examples see [Röhmel and Mansmann, 1999b, p. 151-153].

The issue of the most appropriate hypotheses is not pursued further, instead a general statistical methodology will be presented which allows in principle to deal with any isotonic curve h.

5.2 Asymptotic theory

5.2.1 The LR test for general hypotheses

In this section the likelihood ratio test for (5.1) will be constructed and it will be shown that for smooth h the asymptotic distribution is $\frac{1}{2}+\frac{1}{2}F_{\chi^2_1}$, exactly as for the case where h is the identity (cf. Robertson and Wright [1981, Th. 4.2]).

Further, it will be shown that uniqueness of the MLE depends heavily on the function h. Even for strictly increasing and smooth h uniqueness cannot be guaranteed, in general. Conditions on the "boundary function" h will be given which guarantee uniqueness of the MLE. This highlights an interesting difference between superiority and non-inferiority trials. In superiority trials often the null hypothesis will be convex which immediately implies uniqueness of the MLE, whereas in non-inferiority trials convexity of H_0 is not the typical case. This will be worked out in detail for the difference h_{DI}, the relative risk h_{RR}, and the odds ratio h_{OR}. In particular, it is possible to give explicit expressions for the MLE, constrained to $\vartheta_1 = h(\vartheta_2)$ in these cases.

Throughout the following let $X_1,\dots,X_{n_1} \overset{i.i.d.}{\sim} Bi\,(1,\vartheta_1)$ and $Y_1,\dots,Y_{n_2} \overset{i.i.d.}{\sim} Bi\,(1,\vartheta_2)$ be two independent Bernoulli samples with failure rates ϑ_1 and ϑ_2, respectively. Hence, the joint likelihood is given as

$$L(\vartheta_1,\vartheta_2) = \binom{n_1}{x}\vartheta_1^{\,x}\,(1-\vartheta_1)^{n_1-x}\binom{n_2}{y}\vartheta_2^{\,y}\,(1-\vartheta_2)^{n_2-y} \quad , \tag{5.3}$$

where $x=\sum_{i=1}^{n_1} x_i$ denotes the number of negative responses in treatment group 1 and $y=\sum_{j=1}^{n_2} y_j$ in control group 2, respectively.

Theorem 2.1 guarantees that the MLE constrained to H_0 in (5.1) can be computed on the set $\{\vartheta_1 = h(\vartheta_2)\}$. The next Lemma gives conditions on h which guarantee the uniqueness of the MLE constrained to H_0. Surprisingly, this is not always the case and counter-examples will be given. Let $\vartheta=(\vartheta_1,\vartheta_2)$, $\Theta=[0,1]^2$. Let $\hat\vartheta=(\hat\vartheta_1,\hat\vartheta_2)=(\frac{x}{n_1},\frac{y}{n_2})$, the unconstrained MLE.

Lemma 5.1 *Let $\Theta_0=\{\vartheta:\ \vartheta_1\geq h\,(\vartheta_2)\}$ and assume $X_1,\dots,X_{n_1}\sim Bi\,(1,\vartheta_1)$ i.i.d. and independently $Y_1,\dots,Y_{n_2}\sim Bi\,(1,\vartheta_2)$ i.i.d., where $n_1,n_2\geq 1$. Let h be continuous and increasing, and not identically 1. Let h be twice differentiable, $h\in C^2[0,1]$. The constrained MLE $\hat\vartheta^*$ is unique, if on the set $\Theta_h=\{\vartheta_2 \mid \vartheta_1=h(\vartheta_2),\vartheta_1\in[0,1]\}$ we have that*

(i) $\quad -(h')^2+h\cdot h''\leq 0 \quad$ *and* $\quad -(h')^2-h''+h\cdot h''\leq 0$,

or if

(ii) h *is convex.*

Proof: Define the function $\Psi(\vartheta_2) = \ell(h(\vartheta_2), \vartheta_2)$, where

$$\ell(\vartheta_1, \vartheta_2) := x \log \vartheta_1 + (n_1 - x) \log(1 - \vartheta_1) + y \log \vartheta_2 + (n_2 - y) \log(1 - \vartheta_2)$$

denotes the log-likelihood, $\ell = \log L$ (omitting the constant term $\log \binom{n_1}{x} + \log \binom{n_2}{y}$). We have

$$\begin{aligned}
\Psi''(\vartheta_2) &= -\frac{y}{\vartheta_2^2} - \frac{n_2 - y}{(1-\vartheta_2)^2} - \frac{x}{h^2(\vartheta_2)}(h'(\vartheta_2))^2 + \frac{x}{h(\vartheta_2)} h''(\vartheta_2) \\
&\quad - \frac{n_1 - x}{(1 - h(\vartheta_2))^2}(h'(\vartheta_2))^2 - \frac{n_1 - x}{1 - h(\vartheta_2)} h''(\vartheta_2) \\
&= -\frac{y}{\vartheta_2^2} - \frac{n_2 - y}{(1-\vartheta_2)^2} + \frac{x}{h^2(\vartheta_2)}\big(- (h'(\vartheta_2))^2 + h(\vartheta_2) \cdot h''(\vartheta_2)\big) \\
&\quad + \frac{n_1 - x}{(1 - h(\vartheta_2))^2}\big(- (h'(\vartheta_2))^2 - (1 - h(\vartheta_2)) h''(\vartheta_2)\big) \ .
\end{aligned}$$

Now, if (i) is fulfilled, $\Psi''(\vartheta_2) < 0$, and hence Ψ is strictly concave on the set Θ_h.

In order to prove (ii), observe that ℓ is strictly concave and hence the maximum on Θ_0 which is convex since h is convex, is unique. □

If Theorem 2.1 is applied to the functions in (5.2), it is found that the MLE can always be computed on the set Θ_h for $h = h_{DI}, h_{RR}, h_{OR}$, respectively, and that the MLE is unique. This follows from Lemma 5.1 for h_{DI} and h_{RR} by (ii), whereas for h_{OR} observe that the left hand side of (i) reads as

$$-(h'(\vartheta_2))^2 + h(\vartheta_2) \cdot h''(\vartheta_2) = -\frac{\theta_0^2}{(1 + \vartheta_2(\theta_0 - 1))^3} \leq 0 \ ,$$

and the right hand side of (i) as

$$-(h'(\vartheta_2))^2 - h''(\vartheta_2) + h(\vartheta_2) \cdot h''(\vartheta_2) = -\frac{\theta_0}{(1 + \vartheta_2(\theta_0 - 1))^3} \leq 0 \ .$$

It is interesting to note that condition (ii) is in general not satisfied by most hypotheses for non-inferiority (see Röhmel and Mansmann [1999b, Fig. 1d-1f]). However, in superiority trials these hypotheses are more important, since here H_0 will be a convex set in many cases. This makes a subtle distinction between non-inferiority and superiority trials: The constrained MLE in the latter case will be typically a projection onto a convex set (the

null hypothesis), and hence unique, in non-inferiority trials often the alternative is a convex set, hence uniqueness has to be checked carefully, e.g. by means of Lemma 5.1 (i). Observe finally that (i) guarantees that the likelihood function is strictly convex on Θ_h which allows a quick computation by the use of any standard maximization routine.

The constrained ML estimators for ϑ_1 and ϑ_2 are calculated as follows (cf. also Miettinen and Nurminen [1985]):

a) The difference h_{DI}:

$$\begin{aligned}\hat{\vartheta}_1^* &= 2\frac{\sqrt{r^2-3s}}{3}\cos\left[\frac{1}{3}\arccos\left(-\frac{\frac{2r^3}{27}-\frac{rs}{3}+t}{2\left(\frac{\sqrt{r^2-3s}}{3}\right)^3}\right)+\frac{4}{3}\pi\right]-\frac{r}{3}\ ,\\ \hat{\vartheta}_2^* &= \hat{\vartheta}_1^*-\theta_0\ , \end{aligned} \tag{5.4}$$

with

$$\begin{aligned} r &= -\frac{x+y+n_1\left(1+2\theta_0\right)+n_2\left(1+\theta_0\right)}{n_1+n_2}\ ,\\ s &= \frac{y+x(1+2\theta_0)+\theta_0(n_2+n_1(1+\theta_0)}{n_1+n_2}\ ,\\ t &= \frac{-x\theta_0\left(1+\theta_0\right)}{n_1+n_2}\ . \end{aligned}$$

b) The relative risk h_{RR}:

$$\begin{aligned}\hat{\vartheta}_2^* &= \frac{1}{2\left(n_1+n_2\right)\theta_0}\Big[x+n_2+y\theta_0+n_1\theta_0\\ &\qquad +\sqrt{\left(x+n_2+y\theta_0+n_1\theta_0\right)^2-4\left(x+y\right)\left(n_1+n_2\right)\theta_0}\Big]\ ,\\ \hat{\vartheta}_1^* &= \theta_0\hat{\vartheta}_2^*\ . \end{aligned} \tag{5.5}$$

c) The odds ratio h_{OR}:

$$\begin{aligned}\hat{\vartheta}_2^* &= \frac{1}{2n_2\left(\theta_0-1\right)}\Big[\theta_0(x+y-n_1)-x-y-n_2\\ &\qquad +\sqrt{\left(x+y+n_2-x\theta_0-y\theta_0+n_1\theta_0\right)^2+4\left(x+y\right)n_2\left(\theta_0-1\right)}\Big]\ ,\\ \hat{\vartheta}_1^* &= \left[1+\theta_0^{-1}(\hat{\vartheta}_2^{*^{-1}}-1)\right]^{-1}\ . \end{aligned}$$

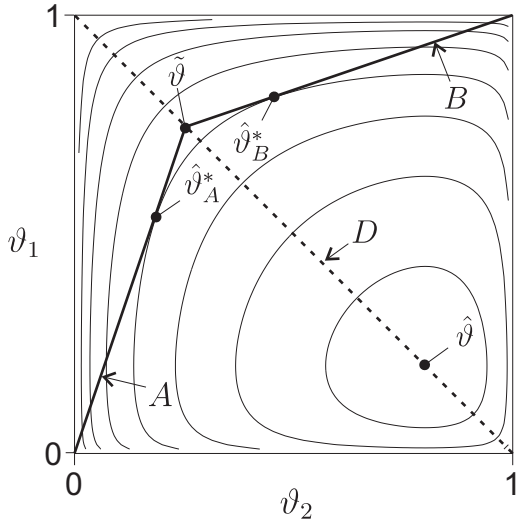

Figure 5.2: Two solutions of the constrained MLE, where $n_1 = n_2 = n$ and $x = n - y$, for the hypotheses in (5.6). Here the contour plot of the two sample binomial likelihood shows the existence of two MLEs $\hat{\vartheta}_A^*$ and $\hat{\vartheta}_B^*$ on each branch A and B of the boundary of the hypothesis H_0, respectively.

These expressions result from straightforward maximization, where the zero of a quadratic (relative risk, odds ratio) or a cubic (difference) polynomial has to be computed.

Remark 5.2 *As mentioned above, uniqueness of the MLE (albeit always located on the set $\{\vartheta : \vartheta_1 = h(\vartheta_2)\}$) is not valid for arbitrary increasing functions h. In fact, various global maxima can occur for certain outcomes (x, y) and hypotheses H_0. A simple class of counter-examples is as follows. Let $n_1 = n_2 = n$ and $\frac{x}{n} = 1 - \frac{y}{n}$, and consider the case where $(\frac{x}{n}, \frac{y}{n}) \in H_1$, i.e. where the constrained MLE does not equal the unconstrained one. Let h be defined as*

$$h(\vartheta_2) = \begin{cases} \frac{1}{\gamma}\ \vartheta_2 & \text{for } \ \vartheta_2 \leq \frac{\gamma}{1+\gamma} \\ \gamma\ \vartheta_2 + 1 - \gamma & \text{else} \end{cases} \tag{5.6}$$

for some constant $0 < \gamma < 1$ (cf. Figure 5.2, here $\gamma = 0.33$). Observe that h is piecewise linear and symmetric (as well as L) with respect to $D = \{(\vartheta_1, \vartheta_2) : \vartheta_1 = 1 - \vartheta_2\}$. For any $\gamma \in (0, 1)$ there are exactly two solutions of the MLE (denoted by $\hat{\vartheta}_A^*$ and $\hat{\vartheta}_B^*$ in Figure 5.2) which are symmetrical w.r.t. D, located on each of the two branches A and B of h in (5.6), respectively.

Proof: For the following arguments it is helpful to consult Figure 5.2. Denote the two branches of h in (5.6) as A and B, respectively. Then, a similar argument as in the proof

of Lemma 5.1 shows that the function $\Psi(\vartheta_2) = \ell(h(\vartheta_2), \vartheta_2)$ (defined in the proof of Lemma 5.1) is strictly concave on each branch A and B, respectively. Hence, on each branch a global maximum exists, and by symmetry it attains the same value. It remains to show that the maximum is not attained on D, i.e. on the intersection $\tilde{\vartheta} = (1-\tilde{\vartheta}_2, \tilde{\vartheta}_2)$ of A and B. Now, by symmetry

$$\text{grad } \ell(1-\tilde{\vartheta}_2, \tilde{\vartheta}_2) = (-a, a)$$

for some value $a > 0$, since ℓ is strictly concave on D and by assumption the maximum which is located in $\hat{\vartheta} \in D$ is in H_1. Further,

$$\Psi'(\tilde{\vartheta}_2) = \text{grad } \ell(1-\tilde{\vartheta}_2, \tilde{\vartheta}_2) \cdot (\alpha^{-1}, 1)^\top = a(1-\alpha^{-1}) < 0 \ .$$

This yields a unique maximum of ℓ restricted to $A \backslash \{\tilde{\vartheta}\}$, and thus it is not attained on D. □

In Figure 5.2 the contour lines of the likelihood are displayed for this particular case and $\gamma = 0.33$. The solutions $\hat{\vartheta}_A^*, \hat{\vartheta}_B^*$ are such that the hypothesis function h is tangent to the likelihood. From this it can also be drawn that various other hypotheses may even lead to more than two solutions of the MLE.

Remark 5.3 *Theorem 2.1 is related to a Theorem of Röhmel and Mansmann [1999b, p. 161], who showed that for fixed $(\overline{x}, \overline{y}) := \hat{\vartheta}$ the supremum over $\vartheta \in \Theta_0$ of $\sum_{T(x,y) \leq T(\overline{x},\overline{y})} L(\vartheta)$ is attained at the boundary of Θ_0, provided that the statistic T satisfies a convexity condition "C" introduced by Barnard [1947]. It states that for any $(x, y) \in CR$ (the critical region of a test) it holds that $(x, y+1) \in CR$ and $(x-1, y) \in CR$. Note, however, that Theorem 2.1 is different and the proof relies essentially on the uniqueness of the unconstrained MLE.*

Theorem 5.4 *Let h be increasing, $h : [0,1] \to [0,1]$, and $h \in C^{(1)}[0,1]$. Then, under the assumption of Lemma 5.1, for $\vartheta_1 = h(\vartheta_2)$ and for any solution $\hat{\vartheta}^*$ it holds that*

$$-2 \ln \lambda \xrightarrow{\mathcal{D}} \tfrac{1}{2} + \tfrac{1}{2} F_{\chi_1^2} \ ,$$

as $\min\{n_1, n_2\} \to \infty$, such that $\frac{n_1}{n_2} \to c \in (0, \infty)$, where λ is the likelihood ratio given in (2.1) and $F_{\chi_1^2}$ denotes the cumulative distribution function of the square of a standard normal random variable.

Proof: First, note that by means of Lemma 5.1 we have

$$\lambda = \begin{cases} 1 & \text{if } \hat{\vartheta} \in \Theta_0 \\ \dfrac{L(\hat{\vartheta}^*)}{L(\hat{\vartheta})} & \text{if } \hat{\vartheta} \notin \Theta_0 \end{cases} . \tag{5.7}$$

Furthermore, for $t \geq 0$,

$$\begin{aligned} P\left(-2\ln\lambda \leq t\right) &= P\left(\{-2\ln\lambda \leq t\} \cap \left\{\hat{\vartheta}_1 \geq h(\hat{\vartheta}_2)\right\}\right) \\ &\quad + P\left(\{-2\ln\lambda \leq t\} \cap \left\{\hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right\}\right) =: I + II\,. \end{aligned}$$

By means of (5.7) we have $\hat{\vartheta}_1 \geq h(\hat{\vartheta}_2) \Leftrightarrow \lambda = 1 \Leftrightarrow -2\ln\lambda = 0$, and hence, if $\vartheta_1 = h\left(\vartheta_2\right)$,

$$I = P\left(\hat{\vartheta}_1 \geq h(\hat{\vartheta}_2)\right) = P\left(\hat{\vartheta}_1 - \vartheta_1 \geq h(\hat{\vartheta}_2) - h\left(\vartheta_2\right)\right) \overset{n_2, n_1 \to \infty}{\longrightarrow} P\left(Z_1 \geq Z_2\right),$$

where Z_1 and Z_2 are independent normal random variables with mean zero and variance $\tau_1 = \vartheta_1\left(1-\vartheta_1\right)$ and $\tau_2 = c\left(h'\left(\vartheta_2\right)\right)^2 \vartheta_2\left(1-\vartheta_2\right)$, respectively [Ferguson, 1996, p. 45, Theorem 7]. Observe that $P\left(Z_1 \geq Z_2\right) = \frac{1}{2}$ always, even if $h'\left(\vartheta_2\right) = 0$. Now,

$$\begin{aligned} II &= P\left(-2\ln\lambda \leq t | \hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right) P\left(\hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right) \\ &= \frac{1}{2} P\left(-2\ln\lambda \leq t | \hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right) + o\left(1\right) \\ &= \frac{1}{2} P\left(-2\ln\lambda \leq t | -2\ln\lambda > 0\right) + o\left(1\right)\ , \end{aligned}$$

and Theorem 2.4 is applied. In order to apply this Theorem, note that $-2\ln\lambda > 0$ ensures that $\hat{\vartheta}^* \in \Theta_h$ and $\hat{\vartheta} \in [0,1]^2 \setminus \Theta_0$. Referring to the notation of Theorem 2.4 (exception: the function h of the theorem is denoted by g and the index n is suppressed here), we obtain

$$\begin{gathered} U(\vartheta) = \left(\tfrac{x}{\vartheta_1} - \tfrac{n_1 - x}{1-\vartheta_1}\ ,\ \tfrac{y}{\vartheta_2} - \tfrac{n_2 - y}{1-\vartheta_2}\right)^\top,\ \Gamma = diag\left(\tfrac{1}{\sqrt{n_1}}\ ,\ \tfrac{1}{\sqrt{n_2}}\right)\ , \\ \Sigma(\vartheta) = B(\vartheta) = diag\left(\tfrac{1}{\vartheta_1(1-\vartheta_1)}\ ,\ \tfrac{1}{\vartheta_2(1-\vartheta_2)}\right)\ , \end{gathered}$$

and for the constrained model

$$\vartheta = (\eta\ ,\ h(\eta))^\top = g(\eta)\ ,\quad (\eta \in (0,1))$$

we have

$$\Gamma^* = \tfrac{1}{\sqrt{n_1}}\ ,\qquad C(\eta) = (\sqrt{c}\ ,\ h'(\eta))^\top \quad (\text{with } c = \frac{n_2}{n_1})\ ,$$

and thus

$$\Sigma^*(\eta) = B^*(\eta) = \frac{c}{\eta(1-\eta)} + \frac{(h'(\eta))^2}{h(\eta)(1-h(\eta))}\ .$$

Hence Theorem 2.4 gives $II = \frac{1}{2} P\left(\chi_1^2 \leq t\right) + o\left(1\right)$. □

Table 5.1: The actual probability (times 100) for $-2\ln(\lambda) > (\frac{1}{2} + \frac{1}{2}F_{\chi^2_1})_{0.95}$.

		exact level		
		difference	relative risk	odds ratio
n_1, n_2	ϑ_2	$\theta_0 = 0.1$	$\theta_0 = 1.5$	$\theta_0 = 1.5$
10,10	0.1	8.93	5.69	6.15
10,25		10.22	9.46	10.25
25,25		5.33	5.59	6.17
25,10		5.27	6.27	7.01
50,50		5.45	6.32	4.4
50,100		5.19	5.22	6.01
100,100		5.22	5.24	4.52
100,50		5.37	4.86	4.31
500,500		5.05	4.97	4.98
10,10	0.4	5.95	4.76	5.72
10,25		5.63	5.45	5.63
25,25		4.46	5.26	4.36
25,10		5.00	5.47	5.00
50,50		4.51	5.16	4.33
50,100		4.84	4.97	4.61
100,100		5.05	4.86	5.05
100,50		4.74	4.91	4.57
500,500		5.18	4.91	5.04

Unfortunately, the approximation using the asymptotics of Theorem 2.4 does not perform very well for small sample sizes.

In Table 5.1 the actual exact levels are drawn for different parameter constellations when using the 95% quantile $(\frac{1}{2} + \frac{1}{2}F_{\chi^2_1})_{0.95}$ of the asymptotic distribution as the critical value for a level 5% test.

From Table 5.1 it can be seen that the nominal level is exceeded up to twice for small sample sizes. As a very rough rule of thumb it can be stated that the asymptotic test can be recommended if $n_1, n_2 \geq 100$, say. Of course, this will depend on the underlying (unknown) response rates. Note that as $\vartheta_1, \vartheta_2 \to 0$, Theorem 5.4 does not hold anymore, instead a Poisson limit is valid. To overcome this drawback in Section 5.3.1, an exact modification of the asymptotic LR test is presented, i.e. a test which keeps its nominal level exactly.

5.2.2 Other asymptotic approaches

Various other methods for the testing problem in (5.1) have been suggested during the last two decades. It is a difficult task to compare all of these methods simultaneously, since many of them are developed from different viewpoints. For example, exact procedures (Section 5.3) aim at keeping the nominal level exactly (see Chan [1998], Röhmel and Mansmann [1999b], Kang and Chen [2000]) whereas asymptotic procedures aim at a maximal power under the constraint of controlling the nominal level at least quite reasonably. A comprehensive numerical comparison of level and power of three asymptotic procedures for the difference was made by Roebruck and Kühn [1995], who came to the conclusion that Farrington and Manning [1990]'s procedure (FM test hereafter) represents a reasonable compromise between a test which keeps its nominal level quite accurately but still has a very good power. It was demonstrated that with respect to this criterion the FM test outperforms the methods by Dunnett and Gent [1977], Blackwelder [1982] and Rodary et al. [1989].

All asymptotic tests for the difference are similar in spirit, since they are based on the asymptotic normality of the test statistic

$$T_{DI} = \frac{\hat{\vartheta}_1 - \hat{\vartheta}_2 - \theta_0}{\sqrt{Var(\hat{\vartheta}_1 - \hat{\vartheta}_2)}}\ , \tag{5.8}$$

where $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$ are the estimated failure rates from independent Bernoulli samples, $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, respectively. Here n_1 and n_2 refer to the number of patients

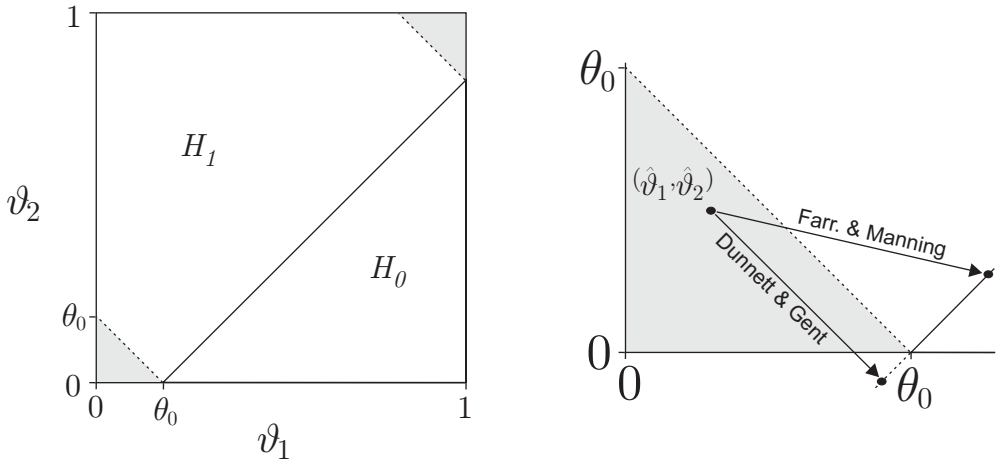

Figure 5.3: Parameter space and hypothesis (left hand side). Projection of the estimators (right hand side).

receiving test treatment 1 and control therapy 2, respectively. The above mentioned approaches differ only in using different estimates for the variance

$$\sigma^2_{DI} := Var\left[\hat{\vartheta}_1 - \hat{\vartheta}_2\right] = \frac{\vartheta_1\,(1-\vartheta_1)}{n_1} + \frac{\vartheta_2\,(1-\vartheta_2)}{n_2}\ . \tag{5.9}$$

To this end Blackwelder [1982] estimated σ^2_{DI} by the unconstrained MLE,

$$\frac{\hat{\vartheta}_1\left(1-\hat{\vartheta}_1\right)}{n_1} + \frac{\hat{\vartheta}_2\left(1-\hat{\vartheta}_2\right)}{n_2}\ ,$$

whereas Dunnett and Gent [1977] and Rodary et al. [1989] suggested to estimate σ^2_{DI} by an MLE constrained to the line $\vartheta_1 = \vartheta_2 + \theta_0$ separating H_0 and H_1, keeping the marginal totals fixed, i.e. $n_1\widetilde{\vartheta}_1 + n_2\widetilde{\vartheta}_2 = n_1\hat{\vartheta}_1 + n_2\hat{\vartheta}_2$. This leads to the estimators

$$\widetilde{\vartheta}_1 = \frac{\hat{\vartheta}_1 + \frac{n_2}{n_1}\left(\hat{\vartheta}_2 + \theta_0\right)}{1+\frac{n_2}{n_1}}\ \ ,\ \widetilde{\vartheta}_2 = \frac{\hat{\vartheta}_1 + \frac{n_2}{n_1}\hat{\vartheta}_2 - \theta_0}{1+\frac{n_2}{n_1}} \tag{5.10}$$

which are inserted in (5.9) instead of ϑ_1 and ϑ_2. In some instances (see grey areas in Figure 5.3) the estimators in (5.10) may fail in that their values may not lie in the range (0,1), respectively. Hence, valid values can be obtained only if the following inequalities are satisfied:

$$\theta_0 \le \hat{\vartheta}_1 + \frac{n_2}{n_1}\hat{\vartheta}_2 \le 1 + \frac{n_2}{n_1}\,(1-\theta_0)\ .$$

Farrington and Manning [1990] circumvent this drawback by using the MLE constrained to $\vartheta_1 = \vartheta_2 + \theta_0$ which strictly lies in the range (0,1). The explicit solution of the maximum likelihood equations is given in (5.4). In order to obtain an estimator of σ^2_{DI}, $\hat{\vartheta}^*_1$ and $\hat{\vartheta}^*_2$ from (5.4) have to be plugged into (5.9) for ϑ_1 and ϑ_2.

For the relative risk and the odds ratio, respectively, asymptotic tests can be constructed analogously to the difference by using score statistics. The resulting test statistic for the relative risk is also introduced by Farrington and Manning [1990] and is given by

$$T_{RR} \;=\; \frac{\hat{\vartheta}_1 - \theta_0\hat{\vartheta}_2}{\sqrt{Var(\hat{\vartheta}_1 - \theta_0^2\hat{\vartheta}_2)}} \,.$$

Analogously to the difference, the MLE constrained to $\vartheta_1 = \theta_0\vartheta_2$ is used to estimate the variance by

$$\hat{\sigma}^2_{RR} := \frac{\hat{\vartheta}_1^*\left(1-\hat{\vartheta}_1^*\right)}{n_1} + \theta_0^2\frac{\hat{\vartheta}_2^*\left(1-\hat{\vartheta}_2^*\right)}{n_2} \,, \tag{5.11}$$

where $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$ are given in (5.5).

If the hypothesis is specified in terms of the odds ratio, a score statistic is given by

$$T_{OR} \;=\; \frac{\log\hat{\theta} - \log\theta_0}{\sqrt{Var(\log\hat{\theta})}} \,,$$

where $\hat{\theta} = x(n_2-y)/(y(n_1-x))$ is the observed odds ratio. The variance of $\log\hat{\theta}$ can be estimated by (cf. Chen et al. [2000])

$$\hat{\sigma}^2_{OR} := \frac{1}{n_1\hat{\vartheta}_1} + \frac{1}{n_1(1-\hat{\vartheta}_1)} + \frac{1}{n_2\hat{\vartheta}_2} + \frac{1}{n_2(1-\hat{\vartheta}_2)} \,.$$

Under the null hypothesis the test statistics T_{RR} and T_{OR} are asymptotically standard normally distributed. Thus, the null hypotheses are rejected if the test statistics are smaller than u_α, respectively.

5.2.3 Level and power comparisons

The asymptotic LR test is compared with the asymptotic score test for h_{DI}, h_{RR} and h_{OR}, respectively. The following parameter settings are included:

- *Equivalence margin*: $\theta_0 \in \{0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2\}$ for h_{DI} and $\theta_0 \in \{1.05, 1.1, 1.25, 1.5, 1.75, 2\}$ for h_{RR} and h_{OR}.
- *Sample sizes*: $n_2 \in \{100, 150, 200, \ldots, 500\}$ and $n_1 \in \{n_2, 1.5\,n_2, 2\,n_2\}$.

This gives 189 different parameter constellations for h_{DI} and 162 constellations for h_{RR} and h_{OR}, respectively. The nuisance parameter ϑ_2 is chosen such that the resulting power

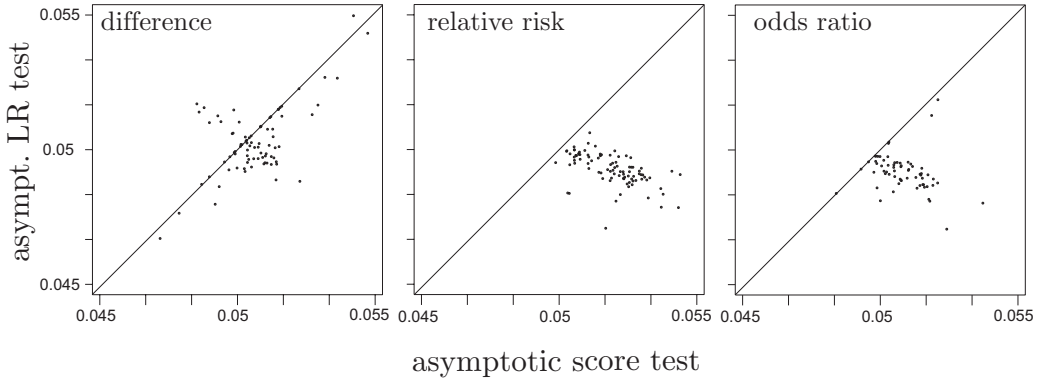

Figure 5.4: The level of the asymptotic 2-sample LR test (vertical axis) in comparison to the asymptotic score test (horizontal axis) for several parameter constellations using the difference, the relative risk and the odds ratio.

for $\vartheta_1 = \vartheta_2$ is in the range $[0.8, 0.9]$, at least for one of the tests compared. If no test leads to a power in this range, the constellation is omitted. Cases of non-feasible settings (i.e. $\vartheta_2 \geq 1 - \theta_0$ for h_{DI}, $\vartheta_2 \geq 1/\theta_0$ for h_{RR}) are omitted, as well. Finally, for h_{DI} 84 parameter constellations were extracted, for h_{RR} 95 parameter constellations and for h_{OR} 64 parameter constellations.

All levels and power values are exactly calculated by computing the exact binomial probabilities given in (5.3) for all (x, y) leading to the rejection of the null hypothesis.

Figure 5.4 shows the level of the asymptotic LR test (vertical axis) and its asymptotic score test (horizontal axes) for the three distance measures difference, relative risk and odds ratio. For the difference the level of the LR test and the score test results between 0.047 and 0.055. The LR test yields slightly smaller levels than its competitor. For the relative risk and the odds ratio the situation is clearer. The level of the LR test almost never exceeds the nominal level 0.05, whereas the level of the score test almost always results in values larger than 0.05. Overall, for sample sizes larger than 100 all tests keep the nominal level sufficiently, ranging between 0.045 and 0.055. It is found that the asymptotic LR test tends to keep the nominal level more accurately.

In Figure 5.5 the power of the asymptotic LR test and its competitors regarding the above mentioned parameter constellations are displayed. Overall, the power differences are minor, in particular for the difference as the distance measure. We find a slight inferiority of the LR test's power for the relative risk and the odds ratio. However, taking into account the more accurate level of the LR test, this inferiority is to be expected. As a conclusion, the LR test is preferred due to its more accurate approximation of the nominal level.

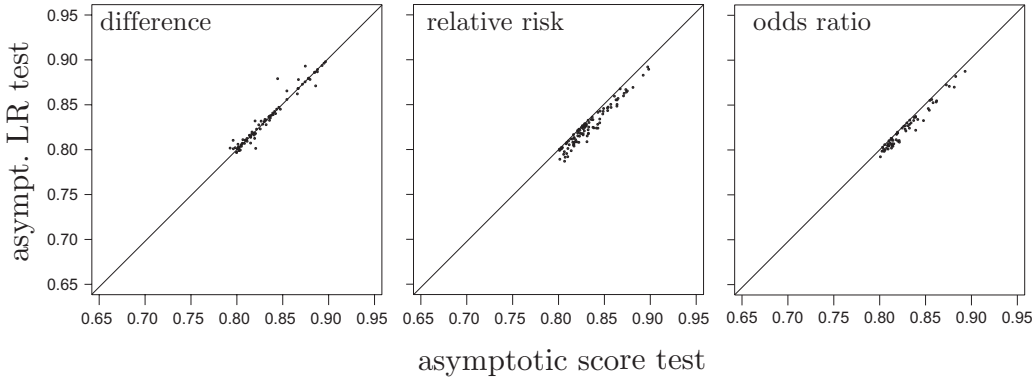

Figure 5.5: The power of the asymptotic 2-sample LR test (vertical axis) in comparison to the asymptotic score test (horizontal axis) for several parameter constellations using the difference, the relative risk and the odds ratio.

5.3 Unconditional exact tests

Exact tests for general hypotheses (5.1) were first introduced in two seminal papers by Barnard [1945, 1947]. It will be shown, however, that Barnard's original test bears intrinsic numerical difficulties due to its specific iterative way to construct the region of rejection. During the last two decades various other exact methods were suggested, most of them were developed for $H_0 \colon \vartheta_1 = \vartheta_2$ (Boschloo [1970], McDonald et al. [1977], Upton [1982], D'Agostino et al. [1988]) or for specific hypotheses in (5.1) (see e.g. Martín Andrés and Silva Mato [1994], Chan [1998]). Finally, Röhmel and Mansmann [1999b] presented a general exact method for arbitrary hypotheses in (5.1), based on ideas of Barnard [1947].

In general, the actual level α^* for a statistical test which specifies the critical region, i.e. the subset CR of the sample space $(0, \ldots, n_1) \times (0, \ldots, n_2)$ for which H_0 is rejected, is calculated by

$$P((X,Y) \in CR \mid (\vartheta_1, \vartheta_2)) = \sum_{(x,y) \in CR} L(\vartheta_1, \vartheta_2) \ , \tag{5.12}$$

where $L(\vartheta_1, \vartheta_2)$ is given in (5.3).

A commonly used approach is to eliminate the unknown parameters ϑ_1 and ϑ_2 by maximizing the function (5.12) over H_0, yielding

$$\alpha^* = \alpha^*(CR) = \max_{\vartheta_1 \geq h(\vartheta_2)} P\left((X,Y) \in CR \mid (\vartheta_1, \vartheta_2)\right). \tag{5.13}$$

Hence, an exact test fulfills $\alpha^* \leq \alpha$, such that equality is attained for some $(\vartheta_1, \vartheta_2) \in H_0$.

5.3.1 The exact LR test

In this section an exact modification of the asymptotic test is given which is based on an idea of Storer and Kim [1990]. This is investigated numerically with various competitors from the literature. It will be shown that the exact (modification of the) LR test in general provides a slightly larger power than its competitors for all specified curves h_{DI}, h_{RR} and h_{OR}.

To guarantee that the LR test keeps its nominal level α, the following modification of the LR statistic λ is applied. In a first step, based on an idea of Storer and Kim [1990], the exact distribution of the LR statistic is estimated by inserting the constrained ML estimates $(\hat{\vartheta}_1^*, \hat{\vartheta}_2^*)$ into (5.3). With that, p-values can be estimated for any outcome (x, y) by calculating

$$p^*(x,y) = \sum_{(a,b):\, \lambda(a,b) \leq \lambda(x,y)} L\left(\hat{\vartheta}_1^*, \hat{\vartheta}_2^*\right) \ , \tag{5.14}$$

where $\lambda(a,b)$ is the likelihood ratio given in (2.1) as a function of the number of failures in both groups.

In a second step these estimated p-values $p^*(x,y)$ are used to sort all possible outcomes $(x_i, y_i) = S_i$ in ascending order. The resulting vector is denoted by

$$S = \left(S_1, \ldots, S_{(n_1+1)\cdot(n_2+1)}\right),$$

and the corresponding increasing values $p^*(S_i) = p_i^*$. Now define

$$\alpha_i^* = \alpha^*\left(\bigcup_{j=1}^{i} S_j\right), \tag{5.15}$$

which denotes the maximal actual level of the rejection region $\bigcup_{j=1}^{i} S_j$ of the "i" smallest values in S with respect to the ordering induced by p^*. Finally, the critical region CR is defined by

$$\arg\max_i \left\{\alpha_i^* \leq \alpha\right\}.$$

Remark 5.5 *Obviously, it is computationally more feasible to calculate the maximum on the boundary of H_0, if possible. Röhmel and Mansmann [1999a] have shown that the maximum is attained always at the boundary $\vartheta_1 = h(\vartheta_2)$, if the test fulfills Barnard's convexity condition "C". Even if we were not able to prove that condition "C" holds for the modified LR test (denoted by exact LR test in the following), the calculation of the maximum in (5.15) can be restricted to the boundary of H_0, since we have checked condition "C" in an extensive numerical investigation and we have found no violation of this condition.*

5.3.2 Other unconditional exact approaches

Barnard's test

Barnard [1947] has constructed an unconditional exact test for $H_0 : \vartheta_1 \geq \vartheta_2$ versus $H_1 : \vartheta_1 < \vartheta_2$. Röhmel and Mansmann [1999b] have recognized that the principle of constructing Barnard's test is directly transferable to the testing problem (5.1). The idea of calculating the critical region is to start with the outcome $(n_2, 0)$, i.e. the most extreme outcome with respect to condition "C" (cf. Remark 5.3). Then the critical region is extended iteratively. Potential next outcomes are the adjacent points $(n_2 - 1, 0)$ and $(n_2, 1)$ which fulfill condition "C". The actual levels α^* (see (5.13)) for $CR = \{(n_2, 0), (n_2 - 1, 0)\}$ and $CR = \{(n_2, 0), (n_2, 1)\}$ are compared. From these adjacent points the outcome is included into the critical region which increases the actual level by the smallest amount. Now, the next adjacent points and their amount to the actual level are calculated to determine the next point to be included in the critical region. This procedure is continued as long as α^* is smaller than the nominal level. Loosely speaking, the critical region is based on the principle to include as much as possible points under the constraint "C" and $\alpha^* \leq \alpha$, where α^* is given in (5.13).

It is found, however, that a serious difficulty encountered with the practical use of Barnard's test consists of the effective numerical computation of its rejection region. To this end the maximum α^* in (5.13) has to be determined numerically for any possible extension of the critical region. Numerically, this can only be achieved by calculating α^* on a discrete grid of the interval $[0, 1]$, the domain of p_2, say. The following algorithm was implemented for computing critical regions:

1) The initial critical region consists of the most extreme possible outcome only: $CR_1 = \{(n_2, 0)\}$.
2) The adjacent outcomes which do not violate condition C are $(n_2 - 1, 0)$ and $(n_2, 1)$. α^* is computed for $CR = CR_1 \cup \{(n_2 - 1, 0)\}$ and $CR = CR_1 \cup \{(n_2, 1)\}$, respectively. The maxima are determined by calculating $P((X, Y) \in CR \mid p_T = p_C + \Delta_0)$ iteratively for $p_C \in \{\epsilon, 2\epsilon, \ldots, 1 - \Delta_0 - 2\epsilon, 1 - \Delta_0 - \epsilon\}$ with, e.g., $\epsilon = 1/1000$. The critical region is extended by the outcome with the smaller α^*. If α^* is (numerically) equal for both outcomes (e.g. if $n_T = n_C$), then $CR_2 = CR_1 \cup \{(n_C - 1, 0), (n_C, 1)\}$.
3) Step 2 is iterated according to condition "C" until α^* exceeds the nominal level.
4) Stop the iteration and choose the preceding critical region.

Note that in step 2 the selection of a possible point to be included in the critical region depends heavily on the grid width ϵ chosen to determine α^*. It is found that this yields

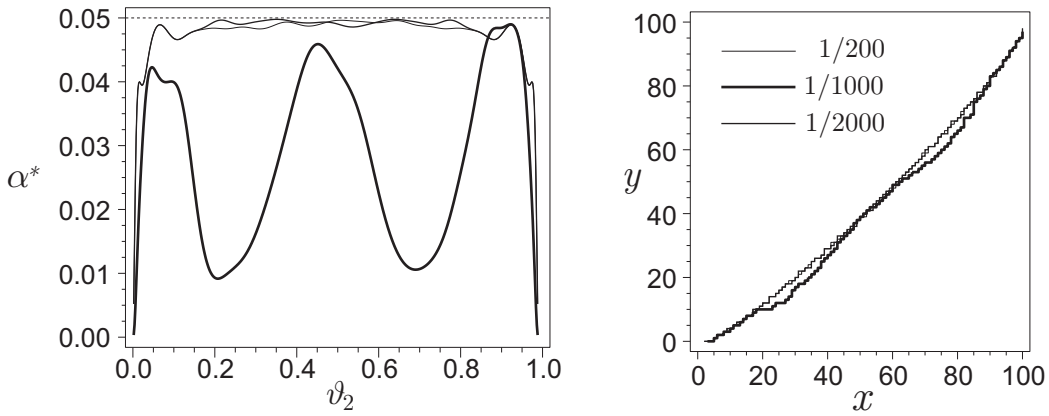

Figure 5.6: Exact level as a function of ϑ_2 (left hand) for calculated rejection regions (right hand) using a grid width of $1/200$ (thin line), $1/1000$ (thick line), and $1/2000$ (medium line).

an intrinsic numerical difficulty, since the corresponding values α^* to be compared are below the numerical precision of any standard software. Due to the iterative structure of the algorithm, a wrong selection of a point in iteration step i will affect the entire subsequent construction and may lead to completely wrong rejection regions. This is in contrast to the subsequent algorithms for the exact LR test, the π_{local} test and Chan's test described below.

Figure 5.6 shows the exact levels as a function of the nuisance parameter ϑ_2 (with $n_1 = n_2 = 100$ and $H_0 : \vartheta_1 \geq \vartheta_2 + 0.01$) on the basis of rejection regions which are calculated using three different grid widths ($1/200$, $1/1000$ and $1/2000$). For a width of $1/1000$, the rejection region is completely degenerated. The right hand figure displays the corresponding rejection regions.

This will be illustrated with the following numerical example. For the construction of the rejection region it is essential that the probabilities corresponding to the potential points are ordered correctly. These probabilities may be extremely small, in particular for larger sample sizes. E.g., for $n_1 = n_2 = 100$, $\theta_0 = 0.01$ and the initial critical region $CR = \{(100, 0)\}$ the maximum is $\alpha^* \approx 8.34 * 10^{-62}$ (for $p_2 = 0.495$). In contrast, for unconditional exact tests which use a test statistic T as an ordering criterion this is a minor problem (just rounding errors may cause difficulties). As a consequence, if T_i and T_j are wrongly ordered in the sequence $T_1, \ldots, T_i, T_j$, this will not affect the subsequent T_k. However, by constructing the critical region of Barnard's test the preceding sequence of the points in the rejection region essentially determines all subsequent points. To illustrate this, assume that the correct "Barnard ordering" with $z_i := (x_i, y_i)$ is $z_1, \ldots, z_i, z_{i+1}$, and

$$\max_{(\vartheta_1,\vartheta_2)\in H_0} P\left((X, Y) \in \{z_1, \ldots, z_i, z_{i+1}\} \mid (\vartheta_1, \vartheta_2)\right) \leq \alpha \ .$$

Assume further, that instead of z_i the outcome z_i' is wrongly inserted into the critical region. Then it may happen that the "correct" outcomes z_i and/or z_{i+1} will not be included into the critical region, since

$$\max_{(\vartheta_1,\vartheta_2)\in H_0} P\left((X,Y) \in \{z_1,\ldots,z_i',z_i\} \mid (\vartheta_1,\vartheta_2)\right) > \alpha \ ,$$

$$\max_{(\vartheta_1,\vartheta_2)\in H_0} P\left((X,Y) \in \{z_1,\ldots,z_i',z_{i+1}\} \mid (\vartheta_1,\vartheta_2)\right) > \alpha \ ,$$

or

$$\max_{(\vartheta_1,\vartheta_2)\in H_0} P\left((X,Y) \in \{z_1,\ldots,z_i',z_i,z_{i+1}\} \mid (\vartheta_1,\vartheta_2)\right) > \alpha \ .$$

Remark 5.6 *There is some different terminology in the literature. In various papers and software packages "Barnard's test" does not refer to the test introduced by Barnard [1947]. E.g., the software product StatXact refers to "Barnard's test" as the unconditional exact test from Chan (see below) in case of the testing problem (5.1). The software product Testimate advertises the "Barnard type exact test for non-inferiority using the Röhmel-Mansmann procedure" which is equal to the π_{local} test (see below).*

The π_{local} test

Röhmel and Mansmann [1999b] have suggested an additional unconditional exact test for the problem (5.1). Here probabilities $\pi_{min}(x,y)$ are calculated for all possible outcomes (x,y) $(0 \le x \le n_1, 0 \le y \le n_2)$ which are denoted by the "smallest possible p-values according to Barnard's condition C". These are the null probabilities for an outcome (i,j) with $i \le x$ and $j \ge y$:

$$\pi_{min}(x,y) = \max_{H_0} \sum_{i\le x} \binom{n_1}{i} \vartheta_1{}^i (1-\vartheta_1)^{n_1-i} \sum_{j\ge y} \binom{n_2}{j} \vartheta_2{}^j (1-\vartheta_2)^{n_2-j} \ .$$

The set of all possible outcomes (x,y) is sorted in ascending order by $\pi_{min}(x,y)$. The test is constructed in the same way as the exact LR test (Section 5.3.1), but it uses $\pi_{min}(x,y)$ as the ordering criterion. This test is applicable for all specifications of a monotone curve h.

Chan's test

In an approach recommended by Chan [1998], the test statistic of Farrington and Manning [1990] (cf. Section 5.2.2) is used to construct an unconditional exact test for h_{DI} and h_{RR}. The critical region of Chan's test is constructed in the same way as the exact

LR test. However, Farrington and Manning's test statistic is used as the ordering criterion. Röhmel and Mansmann [1999a] have remarked that searching for the maximum at the boundary of H_0 might be not correct here, since Chan did not prove that his ordering criterion fulfills the condition "C"- in contrast to Barnard's test and the π_{local} test. However, Chan (author's reply) has given a heuristic argument that the condition is satisfied. In the meantime Martín Andrés and Herranz Tejedor [2003] have given a rigorous proof for this statement.

Fisher's exact unconditional test

McDonald et al. [1977] have suggested to use the conditional exact p-values of Fisher's exact test ($x + y$ is fixed, see e.g. Gart [1971]) as the ordering criterion to construct an unconditional exact version for the classical null hypothesis $H_0 \colon \vartheta_1 = \vartheta_2$. For shifted hypotheses $H_0 \colon \vartheta_1 = h_{OR}(\vartheta_2)$, the conditional version (cf. Lehmann [1986, Chapter 4.5]) using the generalized hypergeometric distribution yields conditional exact p-values:

$$p_C^* := P(X \leq x \mid X + Y = t) = \frac{\sum_{i=0}^{x} \binom{n_1+n_2}{t-i} \binom{n_1}{i} \theta_0^i}{\sum_{j=0}^{t} \binom{n_1+n_2}{t-j} \binom{n_1}{j} \theta_0^j} \; .$$

An unconditional exact version is constructed in the same way as before using the conditional exact p-values p_C^* as the ordering criterion. This adaptation, denoted by *Fisher's exact unconditional test* in the following, is included in the comparison using the odds ratio.

5.3.3 Power investigation

The exact LR test is compared with the above described unconditional exact approaches. For the difference and the relative risk Chan's approach and the π_{local} test are contrasted with the exact LR test.

For the odds ratio, only the π_{local} test and Fisher's exact unconditional test are applicable from the above described approaches. The power properties of these tests are not documented in the literature so far, hence this will be done in the following.

Remark 5.7 *The comparison with Barnard's test (more precise: with Röhmel and Mansmann's adaptation of Barnard's test for the hypothesis (5.1)) is omitted since the investigations in Section 5.3.2 have shown that this test is not applicable in practice due to intrinsic numerical difficulties.*

All tests under investigation are exact methods, i.e. they all keep the nominal level exactly. In the following, these tests are compared numerically for the three distance

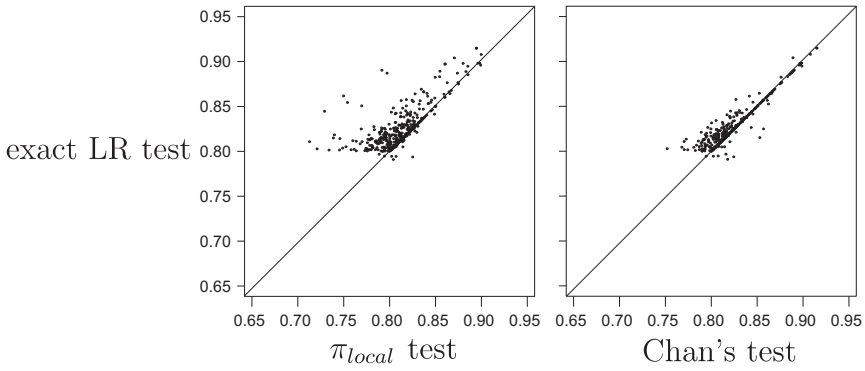

Figure 5.7: The power of the exact LR test (vertical axis) in comparison to the π_{local} test and Chan's test (horizontal axis) for several parameter constellations with $h_{DI}(\vartheta_2) = \vartheta_2 + \theta$.

measures difference, relative risk and odds ratio w.r.t. power for a broad scenario of parameter settings $(\theta_0, n_1, n_2, \vartheta_2)$:

- *Equivalence margin*: $\theta_0 \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$ for h_{DI} and $\theta_0 \in \{1.1, 1.25, 1.5, 2, 2.5\}$ for h_{RR} and h_{OR}.
- *Sample size*: Balanced sample sizes $n_1 = n_2 \in \{20, 25, 30, 35, 40, 50, 60, 80, 100\}$ and unbalanced sample sizes $(n_1, n_2) \in \{(30, 20), (40, 20), (50, 25), (60, 30), (60, 40), (80, 40), (80, 50), (80, 60), (100, 50), (100, 60), (100, 80)\}$.
- *Nuisance parameter*: $\vartheta_2 \in \{0.1, 0.2, 0.3, 0.5, 0.8, 0.9\}$.

This gives 600 different parameter configurations for every function h. Configurations regarding the difference and the relative risk are omitted in case of non-feasible settings (i.e. $\vartheta_2 \geq 1 - \theta_0$ for h_{DI}, $\vartheta_2 \geq 1/\theta_0$ for h_{RR}). The parameters $\theta_{DI} \leq 0$, $\theta_{RR} \leq 1$, and $\theta_{OR} \leq 1$ are chosen such that the resulting power is larger than 0.8, at least for one of the tests compared. Of course, for small sample sizes and small θ_0 there exist parameter constellations, for which no test achieves a power larger than 0.8. On the other hand, for large sample sizes and large θ_0 some parameter constellations result in a power larger than 0.9 for all tests. These cases are omitted, too. Finally, for h_{DI} 410 parameter constellations were extracted, for h_{RR} 330, and for h_{OR} 522. The resulting values of the power function are calculated exactly for all tests under investigation by computing the exact binomial probabilities given in (5.3) for all $(x, y) \in CR$.

The Figures 5.7, 5.8 and 5.9 show the power of the exact LR test (vertical axes) and its competitors (horizontal axes) for the three distance measures h_{DI}, h_{RR} and h_{OR}, respectively.

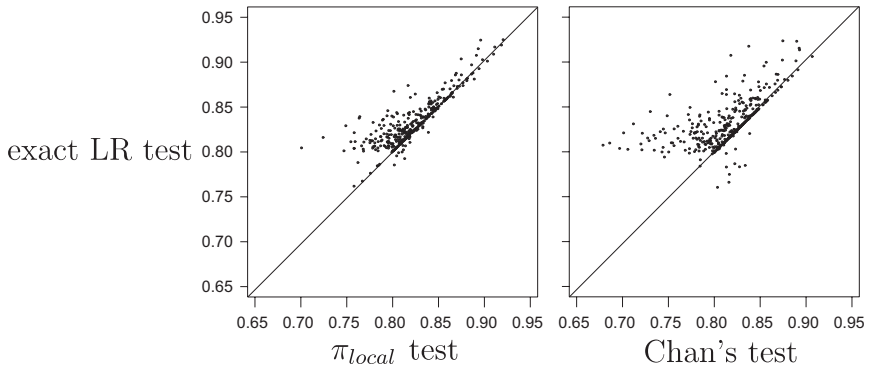

Figure 5.8: The power of the exact LR test (vertical axis) in comparison to the π_{local} test and Chan's test (horizontal axis) for several parameter constellations with $h_{RR}(\vartheta_2) = \vartheta_2\ \theta$.

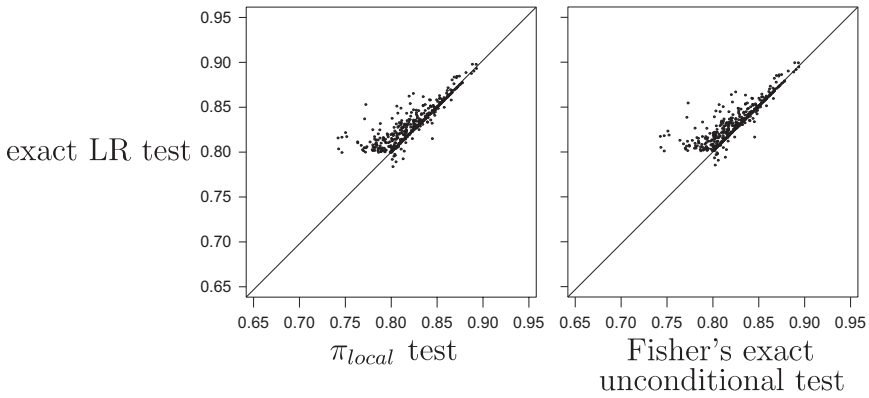

Figure 5.9: The power of the exact LR test (vertical axis) in comparison to the π_{local} test and Fisher's exact unconditional test (horizontal axis) for several parameter constellations with $h_{OR}(\vartheta_2) = \frac{\theta}{\theta+\vartheta_2^{-1}-1}$.

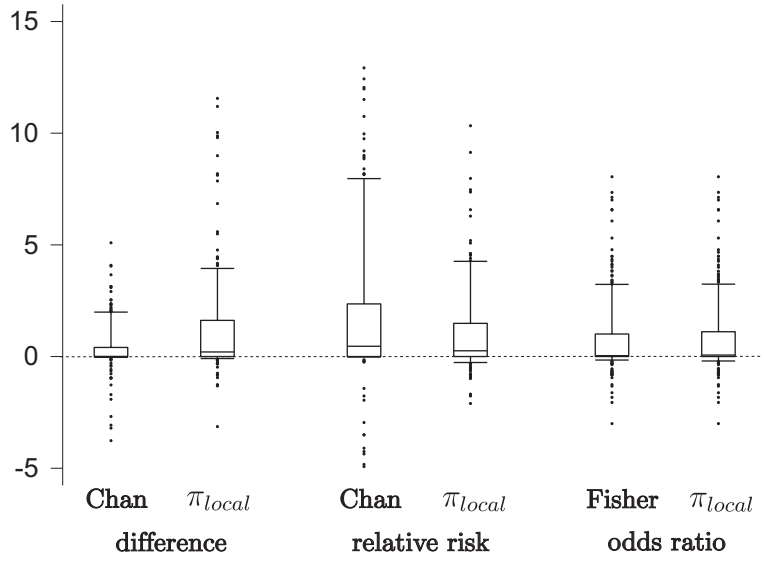

Figure 5.10: Boxplot (whiskers are the 5% and 95% quantiles) for the power differences (times 100) between the exact LR test and its competitors for the three distance measures difference, relative risk and odds ratio.

It is found that in general the power differences between the exact LR test and its competitors are small. Nevertheless, for all three distance measures the power of the exact LR test tends to be larger. In some cases the power enhancement is up to 0.1, whereas the inferiority is much smaller, if present at all. In order to illustrate this, the differences of the exact LR test's power and the power of its competitors are displayed in Figure 5.10 (results of all distance measures combined). The LR test performs better in most cases of parameter constellations, but the median power enhancement is always near zero.

The most extreme power differences and its parameter constellations are displayed in the Appendix (Tables B.1, B.2 and B.3). Those values are displayed separately, where the power of the exact LR test differs from the largest power of its competitors by more than 0.015 (for h_{DI} or h_{RR}) and 0.03 (for h_{OR}), respectively.

The computational time in order to compute the critical regions of all tests is different. It is found that Chan's test is the fastest method. The other methods are more time consuming: Fisher's exact unconditional test requires about 1.5 times of the computational time of Chan's test, the π_{local} test 4 times and the exact LR test about 6 times. In summary, however, all tests under consideration are computationally feasible and numerically stable.

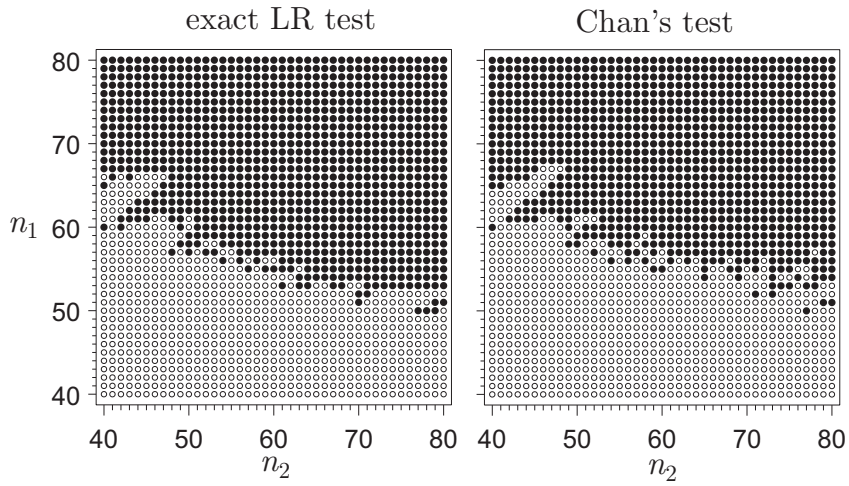

Figure 5.11: Visualization of sample sizes (n_1, n_2), for which the power is larger (black dots) or less (white dots) than 0.8, specifying $\theta_0 = 0.15$, $\alpha = 0.05$ and $p_2 = 0.1$.

5.3.4 Sample size determination

In this section we briefly discuss various issues encountered with the sample size determination when planning a non-inferiority trial in order to control the type II error. One might think at a first glance that this will be in general achieved when the sample sizes n_1 and n_2 in both groups are equal. This is, however, not true (e.g. mentioned by Farrington and Manning [1990], Blackwelder [1993]) when the total number of observations $n_1 + n_2$ is kept fixed.

In order to illustrate this effect, in Figure 5.11 the allocations of sample sizes where $40 \leq n_1, n_2 \leq 80$ are displayed for the LR test and Chan's test, specifying $\theta_0 = 0.15$, $\alpha = 0.05$ and $p_2 = 0.1$. The black dots indicate an allocation of sample sizes for which the test results in a power larger than 80%. For the white dots the power is less than 80%. From this figure the following conclusions can be drawn. The choice of equal sample sizes (displayed on the diagonal $n_1 = n_2$) is not optimal in the sense that the total sample size can be reduced for a different allocation.

In fact, Figure 5.11 shows that the total sample size can be reduced by overweighting group 1 (i.e. $n_1/n_2 > 1$). This was found for various other values of p_2. We need 56 patients per group for a balanced allocation in order to achieve a power of 80%. Minimizing the total number of observations, where the power of 80% is kept fixed, gives in this particular case for the exact LR test and Chan's test the same result, $(n_1, n_2) = (60, 40)$. Hence, the total sample size can be reduced by 12, i.e. by about 10% of the total sample size using a balanced design.

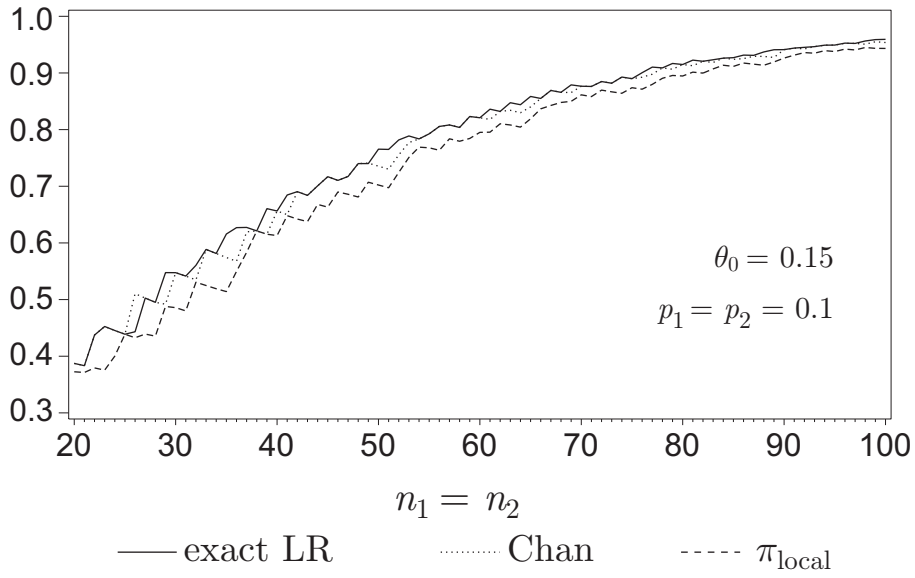

Figure 5.12: Exact power of the exact LR test, Chan's test and the π_{local} test as a function of the sample size (balanced).

As described in Skipka and Trampisch [2001] and Finner and Strassburger [2001], exact tests do not have a monotone increasing power function, in general. In particular, Figure 5.11 shows that for all tests there are pairs of (n_1, n_2), for which the power is larger than for $n_1 + 1$ or $n_2 + 1$. This is found for all tests under investigation. Figure 5.12 shows the exact power as a function of $n_1 = n_2$.

Due to the lack of monotonicity of the power function of these tests, it is computationally extremely intensive to determine the optimal allocation of sample size. A way out of this problem might consist in asymptotic considerations. However, we will not pursue this topic here and leave it as a challenging task for further research.

5.4 Examples

Heliobacter pylori: In a multicenter randomized double-blind study in Heliobacter pylori-positive patients, Dammann et al. [2000] compared the eradication rate of two pantoprazole-based triple therapies of different length. One group (PCM-7) received a combination of pantoprazole, clarithromycin and metronidazole during the first 7 days, followed by 7 days with placebo tablets. The other group (PCM-14) received the same combination of drugs for 14 days. An equivalence margin for the odds ratio of $\theta_0 = 0.33$ was specified. For the intention-to-treat (ITT) population, eradication rates of 89/121 for PCM-7 and 92/123 for PCM-14 were obtained. In the notation of the previous sections (referring to failure rates), this results in $\hat{\vartheta}_1 = 32/121$ and $\hat{\vartheta}_2 = 31/123$. All tests

described above show the non-inferiority of PCM-7 as compared to PCM-14 for the odds ratio, with $\theta_0 = 3.03 \approx 1/0.33$ (p-values: 0.00021 with exact LR test, 0.00025 with the π_{local} test and Fisher's exact unconditional test, respectively). The exact LR test gives a slightly smaller p-value than its competitors.

The corresponding test-based upper 95%-confidence limits for the odds ratio (the smallest θ_0, for which the p-value is smaller than 0.05) are 1.76 (exact LR test) and 1.74 (π_{local} and Fisher's exact unconditional test). Interestingly, the confidence limit based on the exact LR test is slightly larger, albeit in general this test was seen to be more powerful than π_{local} and Fisher's exact unconditional version. Dammann et al. [2000] calculated a lower 95% confidence limit (for the eradication rates) of 0.579, based on the Mantel-Haenszel test - presumably stratified for centers (it is not exactly described in their paper). In the previously used notation this results in $1/0.579 \approx 1.73$ which is similar to our findings, altogether.

Human scabies: In a randomized controlled clinical trial Chouela et al. [1999] compared the therapeutic equivalence of ivermectin (an antihelmintic agent) and lindane (control) for the treatment of human scabies. The sample size was 43; 19 patients received ivermectin and 24 patients received lindane. Chouela et al. fixed the equivalence margin to 0.2 for the difference and argued that ivermectin is much simpler applicable than lindane. It is drawn from Chouela et al. that 29 days after the treatments were administered, 18 of 19 patients treated with ivermectin (5.3% failure rate) and 23 of 24 patients who received lindane (4.2% failure rate) had healing of their scabies. The statistical analysis was performed using Blackwelder's asymptotic test with $\alpha = 0.05$. The p-value was found to be 0.002, hence therapeutic equivalence of ivermectin and lindane was claimed.

In Figure 5.13 the exact levels of the tests of Blackwelder, Chan, and the exact likelihood ratio test are displayed for this example. Figure 5.13 shows that the actual level of Blackwelder's test is heavily exceeded for small and large values of ϑ_2 (up to twice of the nominal level), hence this test is not appropriate here. For example, if the observed failure rate 0.042 is equated with the exact one, the actual level of Blackwelder's test is to be expected as 0.09. The exact LR test has a maximum actual level of 0.049. Finally, the p-value of the data in Chouela et al. for the exact LR test is 0.0087. The p-values for the π_{local} test (0.0152) and Chan's test (0.0172) are somewhat larger. In summary, all of these tests significantly show the therapeutic equivalence of ivermectin and lindane.

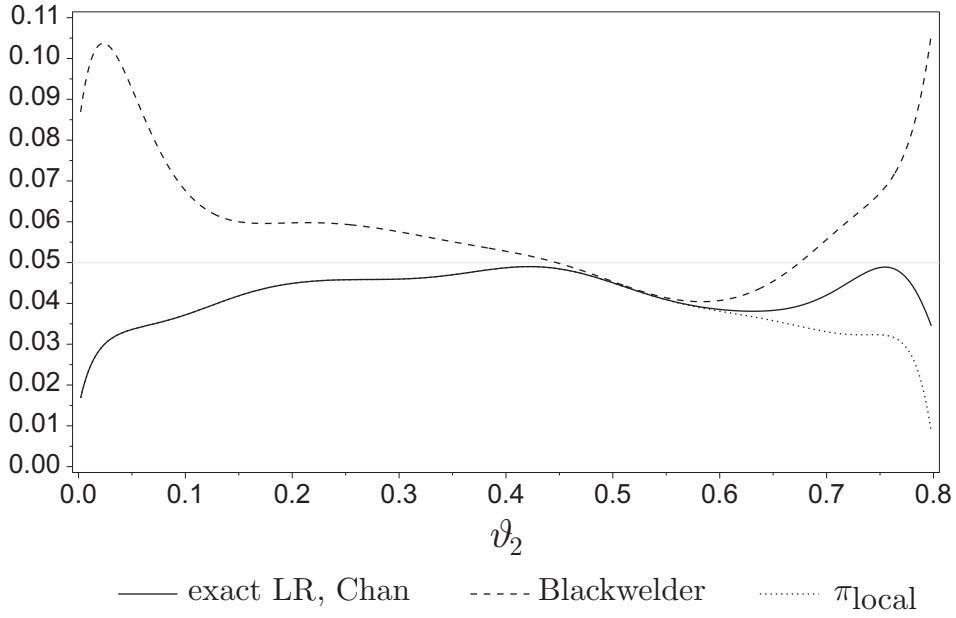

Figure 5.13: Exact level of different statistical tests as a function of ϑ_2 for the parameter constellation of the example "Human scabies".

Remark 5.8 *The stronger condition $\theta_0 = 0.15$ may also be imposed in order to demonstrate therapeutic equivalence. The corresponding p-values are 0.0309 for the exact LR test, 0.04 for Chan's test, and 0.0434 for the π_{local} test. Even if the equivalence margin is chosen smaller, $\theta_0 = 0.13$, say, the exact LR test gives a significant ($\alpha = 0.05$) result (p-value $= 0.0493$). However, Chan's test (p-value $= 0.0544$) and the π_{local} test (p-value $= 0.0677$) do not yield equivalence for $\theta_0 = 0.13$.*

6 Three binomial samples

As seen in Chapter 1, it is often necessary to extend the comparison of a new treatment and a control by a third group. The third group can be a placebo to ensure the assay sensitivity of a clinical trial. Other three-armed clinical trials aim to show that a new treatment is relevantly superior or non-inferior with respect to two standard treatments. Furthermore, if a new treatment is applied using two different formulations or doses, three groups are required in order to compare the new treatment to a control. In this chapter, the LR tests for the hypotheses a) - c) described in the introduction are derived. Analogously to the two-sample case, the methodology is given for hypotheses using general functions h (cf. Section 5.1). Power comparisons based on the commonly used distance measures for binomial distributions (difference of rates, relative risk and odds ratio) are carried out for asymptotic and unconditional exact approaches.

6.1 Model and hypotheses

Let $X_{ij} \sim Bi(1, \vartheta_i)$ be three independent Bernoulli samples with failure rates ϑ_i and sample sizes n_i ($j = 1, \ldots, n_i$, $i = 1, 2, 3$). The following null hypotheses are investigated (analogously to the hypotheses a) - c) of Chapter 1):

$$\begin{aligned} H_0^a &: \vartheta_3 \geq h_1(\vartheta_1) \ \vee \ \vartheta_3 \geq h_2(\vartheta_2) \ , \\ H_0^b &: \vartheta_3 \geq h_1(\vartheta_1) \ \wedge \ \vartheta_3 \geq h_2(\vartheta_2) \ , \\ H_0^c &: \vartheta_3 \geq h_1(\vartheta_1) \ \vee \ \vartheta_1 \geq h_2(\vartheta_2) \ . \end{aligned} \tag{6.1}$$

The functions h_1 and h_2 in (6.1) are specified as for the two-sample case (cf. Chapter 5), i.e. both functions are twice differentiable and strict isotonic.

Figure 6.1 shows the null space H_0^a and H_0^b for the three commonly used distance measures (for definition see (5.2)) difference, relative risk and odds ratio. The null space H_0^c is omitted in the figure, since it is H_0^a with interchanged axes.

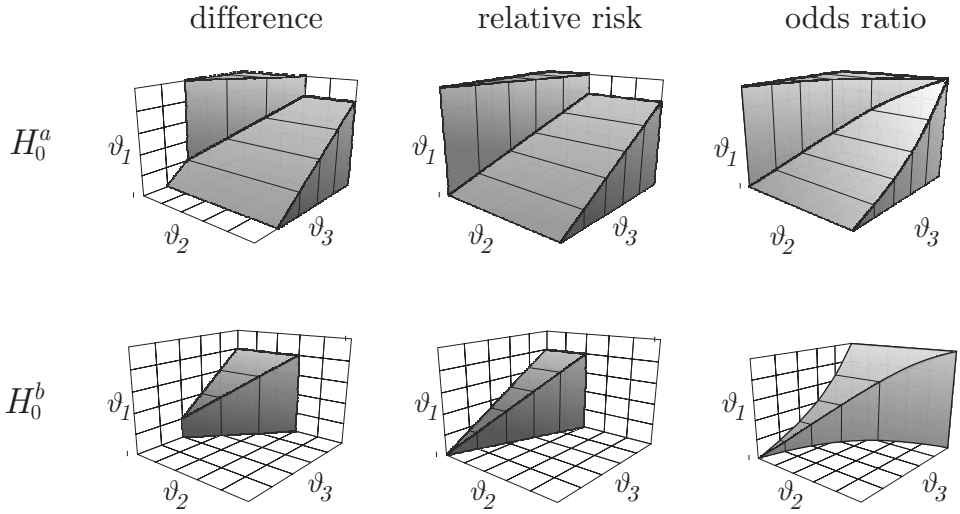

Figure 6.1: Null space for H_0^a and H_0^b .

6.2 Likelihood ratio statistics and asymptotic distribution

In this section the LR statistics and their asymptotic distribution are derived for each of the problems in (6.1). It will be seen that the LR principle for the hypotheses H_0^a and H_0^c leads to the IUT using the pairwise two-sample tests. For H_0^b it will be found that the LR approach is more complicated, since the roots of a 5-degree polynomial have to be calculated.

Let $x_i = \sum_{j=1}^{n_i} x_{ij}$, $i = 1, 2, 3$. Then, the likelihood function is found to be

$$L(\vartheta) = \prod_{i=1}^{3} \binom{n_i}{x_i} \vartheta_i^{x_i} \, (1 - \vartheta_i)^{n_i - x_i} \quad , \tag{6.2}$$

where $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3)^\top$. The unconstrained MLE is given by $\hat{\vartheta} = (x_i/n_i)_{i=1,2,3}$, whereas the MLE constrained to H_0 is given by

$$\hat{\vartheta}^* = \arg\max_{\vartheta \in H_0} L(\vartheta) \ .$$

Thus, the transformed LR $-2 \ln \lambda$ is calculated by

$$T = T(\hat{\vartheta}) := 2[\log L(\hat{\vartheta}) - \log L(\hat{\vartheta}^*)] \ . \tag{6.3}$$

We will show in the next sections that for the hypotheses H_0^a and H_0^c the LR-statistic reduces to a function of the corresponding pairwise two-sample LR-statistics. Therefore, the constrained two-sample MLEs introduced in Chapter 5.2.1 are required. Throughout the following the constrained two-sample MLEs are denoted by

$$\hat{\vartheta}^*_{n_1,n_2,x_1,x_2,h} := \arg\max_{\vartheta} \vartheta^{x_1}(1-\vartheta)^{n_1-x_1}(h(\vartheta))^{x_2}(1-h(\vartheta))^{n_2-x_2} \ . \tag{6.4}$$

Hypotheses H_0^a and H_0^c

Here we derive the test statistic and its asymptotic null distribution for

$$H_0^a : \vartheta_3 \geq h_1(\vartheta_1) \ \vee \ \vartheta_3 \geq h_2(\vartheta_2) \quad \text{vs.} \quad \neg H_0^a \ .$$

Obviously, if $\hat{\vartheta} \in H_0^a$, the LR-statistic equals zero. Thus, let $\hat{\vartheta} \notin H_0^a$. The boundaries of the pairwise null spaces are denoted by

$$\begin{aligned} S_1 &:= \{\vartheta \in H_0^a | \vartheta_3 = h_1(\vartheta_1)\} \ , \\ S_2 &:= \{\vartheta \in H_0^a | \vartheta_3 = h_2(\vartheta_2)\} \ . \end{aligned} \tag{6.5}$$

Since

$$\max_{\vartheta \in H_0^a} L(\vartheta) \overset{Th.\ 2.1}{=} \max_{\vartheta \in \partial H_0^a} L(\vartheta) \leq \max_{S_1 \cup S_2} L(\vartheta) \leq \max_{\vartheta \in H_0^a} L(\vartheta) \ ,$$

the maximum over H_0^a is calculated by

$$\max_{\vartheta \in \partial H_0^a} L(\vartheta) = \max_{S_1 \cup S_2} L(\vartheta) = \max\{\max_{S_1} L(\vartheta), \max_{S_2} L(\vartheta)\} \ .$$

The parameter ϑ_2 is unconstrained in S_1 and ϑ_1 is unconstrained in S_2. Thus, with (6.4) the MLE constrained to H_0^a is one of the constrained two-sample MLEs:

$$\begin{aligned} \hat{\vartheta}^*_{S_1} &:= (\hat{\vartheta}^*_{n_1,n_3,x_1,x_3,h_1} \ , \ \hat{\vartheta}_2 \ , \ h_1(\hat{\vartheta}^*_{n_1,n_3,x_1,x_3,h_1}))^\top \ , \\ \hat{\vartheta}^*_{S_2} &:= (\hat{\vartheta}_1 \ , \ \hat{\vartheta}^*_{n_2,n_3,x_2,x_3,h_2} \ , \ h_2(\hat{\vartheta}^*_{n_2,n_3,x_2,x_3,h_2}))^\top \ . \end{aligned}$$

Therefore, the test statistic (6.3) is given by

$$T = 2[\ln L(\hat{\vartheta}) - \ln \max\{L(\hat{\vartheta}^*_{S_1}), L(\hat{\vartheta}^*_{S_2})\}] = \min\{T_1, T_2\} \ ,$$

where

$$T_i = 2[\ln L(\hat{\vartheta}) - \ln L(\hat{\vartheta}^*_{S_i})] \ , \ i = 1, 2 \ . \tag{6.6}$$

Hence, the test statistic T is equal to the two-sample test for the hypothesis $H_0 : \vartheta_3 \geq h_1(\vartheta_1)$ in case of $L(\hat{\vartheta}^*_{S_1}) \geq L(\hat{\vartheta}^*_{S_2})$, otherwise it is equal to the two-sample test for the hypothesis $H_0 : \vartheta_3 \geq h_2(\vartheta_2)$.

The following theorem guarantees that, asymptotically, H_0^a is rejected at level α if T is larger than the $(1-\alpha)$-quantile of the distribution of $\frac{1}{2} + \frac{1}{2}F_{\chi^2_1}$.

Theorem 6.1 *Let $t > 0$. Then, for all $\vartheta \in \partial H_0^a$ it holds that*

$$P(T > t) \leq P(Z > t) + o(1) \ , \tag{6.7}$$

where Z is distributed as $\frac{1}{2} + \frac{1}{2}F_{\chi^2_1}$. Furthermore, for some $\vartheta \in \partial H_0^a$ we have strict equality in (6.7).

Proof: First note that

$$\begin{aligned} P(T > t) &= P(T_1 > t \ , \ L(\hat{\vartheta}^*_{S_1}) \geq L(\hat{\vartheta}^*_{S_2})) \\ &\quad + P(T_2 > t \ , \ L(\hat{\vartheta}^*_{S_1}) < L(\hat{\vartheta}^*_{S_2})) \ . \end{aligned}$$

In case of $\vartheta_3 = h_1(\vartheta_1), \vartheta_3 < h_2(\vartheta_2)$,

$$P(L(\hat{\vartheta}^*_{S_1}) \geq L(\hat{\vartheta}^*_{S_2})) = 1 + o(1) \ ,$$

and thus,

$$P(T > t) = P(T_1 > t) + o(1) \ .$$

Analogously, in case of $\vartheta_3 = h_2(\vartheta_2), \vartheta_3 < h_1(\vartheta_1)$,

$$P(T > t) = P(T_2 > t) + o(1) \ .$$

Hence, if ϑ is not located on the edge $S_1 \cap S_2$, the test statistic follows the same distribution ($\frac{1}{2} + \frac{1}{2}\chi^2_1$) as in the two-sample case (see. Theorem 5.4).

If ϑ is located on the edge $\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)$,

$$P(T > t) = P(T_1 > t \ , \ T_2 > t) \leq P(T_1 > t) \ ,$$

i.e. $P(T > t) \leq \alpha$ for $t = (\frac{1}{2} + \frac{1}{2}F_{\chi^2_1})_{1-\alpha}$ □

Remark 6.2 *Theorem 6.1 shows that the LR test for H_0^a is equal to the IUT, i.e. H_0^a is rejected, if both of the pairwise two-sample LR tests are rejected at level α.*

Following the same arguments as before, the test statistic and its null distribution for

$$H_0^c : \vartheta_3 \geq h_1(\vartheta_1) \ \vee \ \vartheta_1 \geq h_2(\vartheta_2) \quad \text{vs.} \quad \neg H_0^c$$

can be derived. The MLE constrained to H_0^c is calculated by using one of the constrained two-sample MLEs (see (6.4)):

$$\begin{aligned} \hat{\vartheta}^*_{S_1} &:= (\hat{\vartheta}^*_{n_1,n_3,x_1,x_3,h_1} \ , \ \hat{\vartheta}_2 \ , \ h_1(\hat{\vartheta}^*_{n_1,n_3,x_1,x_3,h_1}))^\top \ , \\ \hat{\vartheta}^*_{S_2} &:= (h_2(\hat{\vartheta}^*_{n_2,n_1,x_2,x_1,h_2}) \ , \ \hat{\vartheta}^*_{n_2,n_1,x_2,x_1,h_2} \ , \ \hat{\vartheta}_3)^\top \ . \end{aligned}$$

Analogously to H_0^a, the test statistic (6.3) is calculated by

$$T = \min\{T_1, T_2\} \ ,$$

where T_i ($i = 1, 2$) is given in (6.6).

Hence, the test statistic T is equal to the two-sample test for the hypothesis $H_0 : \vartheta_3 \geq h_1(\vartheta_1)$ in case of $L(\hat{\vartheta}^*_{S_1}) \geq L(\hat{\vartheta}^*_{S_2})$, otherwise it is equal to the two-sample test for the hypothesis $H_0 : \vartheta_1 \geq h_2(\vartheta_2)$.

Asymptotically, H_0^c is rejected if T is larger than the $(1-\alpha)$-quantile of the distribution of $\frac{1}{2} + \frac{1}{2}F_{\chi^2_1}$ which can be proven in the same way as Theorem 6.1.

Hypothesis H_0^b

For hypothesis H_0^b the situation is different, i.e. the LR test is not a combination of two pairwise comparisons. The boundary of the hypotheses

$$H_0^b : \vartheta_3 \geq h_1(\vartheta_1) \wedge \vartheta_3 \geq h_2(\vartheta_2) \quad \text{vs.} \quad \neg H_0^b$$

is given by the union of the surfaces

$$\begin{aligned} K_1 &:= \{\vartheta \in [0,1]^3 \mid \vartheta_3 = h_1(\vartheta_1) \ \wedge \ \vartheta_3 \geq h_2(\vartheta_2)\} \ , \\ K_2 &:= \{\vartheta \in [0,1]^3 \mid \vartheta_3 = h_2(\vartheta_2) \ \wedge \ \vartheta_3 \geq h_1(\vartheta_1)\} \ . \end{aligned}$$

In contrast to the hypothesis H_0^a, the calculation of the LR-statistic cannot be reduced to the two-sample case in general, since K_1 and K_2 are proper subsets of S_1 and S_2 (cf. (6.5)), respectively. Thus, the MLEs constrained to S_1 and S_2 are not included in H_0^b for some outcomes. In that case the MLEs constrained to H_0^b result at the "edge" of H_0^b, i.e. at $K_3 := K_1 \cap K_2$.

Theorem 6.3 *Let $\hat{\vartheta} \notin H_0^b$. With the notation of Section 6.2 the constrained MLE for the hypothesis H_0^b is given by*

$$\hat{\vartheta}^* := \begin{cases} \hat{\vartheta}^*_{S_1} & \\ \hat{\vartheta}^*_{S_2} & \textit{if} \quad \begin{array}{l} \hat{\vartheta}^*_{S_1} \in H_0^b, \ \hat{\vartheta}^*_{S_2} \notin H_0^b \ \vee \ (\hat{\vartheta}^*_{S_2} \in H_0^b, \ T_1 \leq T_2) \\ \hat{\vartheta}^*_{S_2} \in H_0^b, \ \hat{\vartheta}^*_{S_1} \notin H_0^b \ \vee \ (\hat{\vartheta}^*_{S_1} \in H_0^b, \ T_1 > T_2) \\ \hat{\vartheta}^*_{S_1} \notin H_0^b, \ \hat{\vartheta}^*_{S_2} \notin H_0^b \end{array} \\ \hat{\vartheta}^*_{K_3} & \end{cases} ,$$

where $\hat{\vartheta}^*_{K_3} := \arg\max_{K_3} L(\vartheta)$.

Proof: The maximum of $L(\vartheta)$ over H_0^b is attained in $\partial H_0^b = K_1 \cup K_2$.

If $\hat{\vartheta}^*_{S_1} \notin H_0^b$, we have that $\arg\max_{K_1} L(\vartheta) \in K_3$, since $L(\vartheta)$ is isotonic in ϑ_2 $(< \hat{\vartheta}_2)$ for fixed parameters $\vartheta_3 = h_1(\vartheta_1)$. Analogously, $\arg\max_{K_2} L(\vartheta) \in K_3$ holds for $\hat{\vartheta}^*_{S_2} \notin H_0^b$. It follows that $\hat{\vartheta}^* = \hat{\vartheta}^*_{K_3}$.

If $\hat{\vartheta}^*_{S_1} \in H_0^b$ and $\hat{\vartheta}^*_{S_2} \notin H_0^b$, it follows that $\arg\max_{K_2} L(\vartheta) \in S_1$, since $K_3 \subset S_1$. If $\hat{\vartheta}^*_{S_1} \in H_0^b$ and $\hat{\vartheta}^*_{S_2} \in H_0^b$, $\max_{\vartheta \in H_0^b} L(\vartheta) = L(\hat{\vartheta}^*_{S_1})$ for $T_1 \leq T_2$. This proves the case $\hat{\vartheta}^* = \hat{\vartheta}^*_{S_1}$, and by symmetry the case $\hat{\vartheta}^* = \hat{\vartheta}^*_{S_2}$, also. □

It is shown in Theorem 6.3 that the calculation of the MLE can be reduced to the two-sample case if $\hat{\vartheta}^*_{S_1} \in H_0^b$ or $\hat{\vartheta}^*_{S_2} \in H_0^b$. Otherwise, $L(\vartheta)$ has to be maximized under the constraint $\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)$. In order to find the $\arg\max$ in this situation, one has to compute the zeros of

$$\frac{x_3}{\vartheta_3} - \frac{n_3 - x_3}{1 - \vartheta_3} + \frac{h_1'(\vartheta_1)(x_1 - n_1 h_1(\vartheta_1))}{h_1(\vartheta_1)(1 - h_1(\vartheta_1))} + \frac{h_2'(\vartheta_2)(x_2 - n_2 h_2(\vartheta_2))}{h_2(\vartheta_2)(1 - h_2(\vartheta_2))} \ ,$$

which are the roots of a 5-degree-polynomial. This can be determined numerically by Newton's method. Note that any zero in the interior of the null hypothesis can be used as a MLE.

The test statistic T is calculated as in (6.3) with $\hat{\vartheta}^*$ given by Theorem 6.3.

The asymptotic distribution of the LR-statistic T for H_0^b is given by the following Theorem.

Theorem 6.4 *Let $\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)$ and $X = (X_1, X_2, X_3)^\top$ a 3-dimensional normally distributed random vector with zero mean and covariance matrix $\Sigma^{-1}(\vartheta)$, where*

$$\Sigma(\vartheta) := diag\left(\frac{1}{\vartheta_1(1-\vartheta_1)}\ ,\ \frac{1}{\vartheta_2(1-\vartheta_2)}\ ,\ \frac{1}{\vartheta_3(1-\vartheta_3)}\right)\ .$$

Let further, for $i = 1, 2$,

$$\begin{aligned}\Sigma_i(\vartheta) &:= diag\left(\tfrac{1}{\vartheta_i(1-\vartheta_i)}\ ,\ \tfrac{1}{\vartheta_3(1-\vartheta_3)}\right)\ ,\\ C_i(\eta_i) &:= (\sqrt{c_i}\ ,\ h_i'(\eta_i))^\top\ ,\\ \Sigma_i^*(\eta_i) &:= \tfrac{c_i}{\eta_i(1-\eta_i)} + \tfrac{h_i'(\eta_i)^2}{h_i(\eta_i)(1-h_i(\eta_i))}\ ,\end{aligned}$$

and

$$\begin{aligned}C(\eta) &:= (\sqrt{c_1}\ ,\ \sqrt{c_2}[h_2^{-1}(h_1(\eta))]'\ ,\ h_1'(\eta))^\top\ ,\\ \Sigma^*(\eta) &:= \tfrac{c_1}{\eta(1-\eta)} + \tfrac{h_1'(\eta)^2}{h_1(\eta)(1-h_1(\eta))} + \tfrac{c_2([h_2^{-1}(h_1(\eta))]')^2}{h_2^{-1}(h_1(\eta))[1-h_2^{-1}(h_1(\eta))]}\ .\end{aligned}$$

Then, as $\min_{i=1,2,3}\{n_i\} \to \infty$, s.t. $\frac{n_i}{n_3} \to c_i \in (0,\infty)$ $(i = 1, 2)$, it holds for $t > 0$ that $P(T > t) \to p_1(t) + p_2(t) + p_3(t)$, where

$$\begin{aligned}p_1(t) \ :=\ & P\left((X_1,X_3)^\top A_1\begin{pmatrix}X_1\\X_3\end{pmatrix} > t\ \cap\ [X_3 < \frac{h_1'(\vartheta_1)}{\sqrt{c_1}}X_1 \cup X_3 < \frac{h_2'(\vartheta_2)}{\sqrt{c_2}}X_2]\right.\\ & \cap\ B_1X_1 \geq \frac{1}{\sqrt{c_2}}X_2 \cap \left[B_2X_2 < \frac{1}{\sqrt{c_1}}X_1\right.\\ & \left.\left.\cup\ \ \{B_2X_2 \geq \frac{1}{\sqrt{c_1}}X_1 \cap (X_1,X_3)^\top A_1\begin{pmatrix}X_1\\X_3\end{pmatrix} \leq (X_2,X_3)^\top A_2\begin{pmatrix}X_2\\X_3\end{pmatrix}\}\right]\right)\ ,\end{aligned}$$

$$\begin{aligned}p_2(t) \ :=\ & P\left((X_2,X_3)^\top A_2\begin{pmatrix}X_2\\X_3\end{pmatrix} > t\ \cap\ [X_3 < \frac{h_2'(\vartheta_2)}{\sqrt{c_2}}X_2 \cup X_3 < \frac{h_1'(\vartheta_1)}{\sqrt{c_1}}X_1]\right.\\ & \cap\ B_2X_2 \geq \frac{1}{\sqrt{c_1}}X_1 \cap \left[B_1X_1 < \frac{1}{\sqrt{c_2}}X_2\right.\\ & \left.\left.\cup\ \ \{B_1X_1 \geq \frac{1}{\sqrt{c_2}}X_2 \cap (X_1,X_3)^\top A_1\begin{pmatrix}X_1\\X_3\end{pmatrix} > (X_2,X_3)^\top A_2\begin{pmatrix}X_2\\X_3\end{pmatrix}\}\right]\right)\ ,\end{aligned}$$

$$p_3(t) \;:=\; P\left(X^\top A_1 X > t \;\cap\; [X_3 < \frac{h_1'(\vartheta_1)}{\sqrt{c_1}} X_1 \cup X_3 < \frac{h_2'(\vartheta_2)}{\sqrt{c_2}} X_2] \right.$$
$$\left. \cap \quad B_1 X_1 < \frac{1}{\sqrt{c_2}} X_2 \cap B_2 X_2 < \frac{1}{\sqrt{c_1}} X_1 \right) \;,$$

with (suppressing the arguments of the functions $\Sigma(\vartheta)$, $\Sigma_i(\vartheta)$, etc.)

$$\begin{aligned} A &= \Sigma - \Sigma C {\Sigma^*}^{-1} C^\top \Sigma \;, \\ A_i &= \Sigma_i - \Sigma_i C_i \Sigma_i^{*^{-1}} C_i^\top \Sigma_i \quad (i = 1, 2) \;, \\ B_i &= {\Sigma^*}^{-1} C_i^\top \Sigma_i \quad (i = 1, 2) \;. \end{aligned}$$

Proof: If $\hat\vartheta^* = \hat\vartheta^*_{S_i}$ $(i = 1, 2)$, i.e. the MLE constrained to H_0^b is in S_i, the test statistic T_i is given by (6.6). Since for $\hat\vartheta^* = \hat\vartheta^*_{K_3}$ the MLE is calculated under the constraint K_3, the test statistic is given by $T_3 := 2[\ln L(\hat\vartheta) - \ln L(\hat\vartheta^*_{K_3})]$.

With Theorem 6.3 it holds for $t > 0$ that

$$\begin{aligned} P(T > t) = {} & P(T_1 > t \cap \hat\vartheta \notin H_0^b \cap \hat\vartheta^*_{S_1} \in H_0^b \cap [\hat\vartheta^*_{S_2} \notin H_0^b \cup \{\hat\vartheta^*_{S_2} \in H_0^b \cap T_1 \le T_2\}]) \\ & + P(T_2 > t \cap \hat\vartheta \notin H_0^b \cap \hat\vartheta^*_{S_2} \in H_0^b \cap [\hat\vartheta^*_{S_1} \notin H_0^b \cup \{\hat\vartheta^*_{S_1} \in H_0^b \cap T_1 > T_2\}]) \\ & + P(T_3 > t \cap \hat\vartheta \notin H_0^b \cap \hat\vartheta^*_{S_1} \notin H_0^b \cap \hat\vartheta^*_{S_2} \notin H_0^b) \;. \end{aligned}$$

From Lemma 2.5 it follows that T_i $(i = 1, 2, 3)$ is asymptotically equivalent to $(\hat X_i, \hat X_3)\, A_i\, (\hat X_i, \hat X_3)^\top$ if $\vartheta = h(\eta) = (\eta_i, h_i(\eta_i))^\top$ for $i = 1, 2$, and to $\hat X^\top A\, \hat X$ if $\vartheta = h(\eta) = (\eta, h_2^{-1}(h_1(\eta)), h_1(\eta))^\top$ for $i = 3$, where

$$\hat X \;=\; (\hat X_1, \hat X_2, \hat X_3)^\top \;=\; (\sqrt{n_j}(\hat\vartheta_j - \vartheta_j))_{j=1,2,3} \;.$$

Since $\hat\vartheta \in H_0^b$ is equivalent to $\hat\vartheta_3 \ge h_1(\hat\vartheta_1) \;\cap\; \hat\vartheta_3 \ge h_2(\hat\vartheta_2)$ and hence to $\sqrt{n_3}(\hat\vartheta_3 - \vartheta_3) \ge \sqrt{\frac{n_1}{c_1}}(h_1(\hat\vartheta_1) - h_1(\vartheta_1)) \;\cap\; \sqrt{n_3}(\hat\vartheta_3 - \vartheta_3) \ge \sqrt{\frac{n_2}{c_2}}(h_2(\hat\vartheta_2) - h_2(\vartheta_2))$, with $h_i(\hat\vartheta_i) - h_1(\vartheta_i) = h_i'(\vartheta_i)(\hat\vartheta_i - \vartheta_i) + o_p(|\hat\vartheta_i - \vartheta_i|)$ as $\min\{n_i\} \to \infty$, it holds that

$$[\hat\vartheta_3 - h_i(\hat\vartheta_i)] \;-\; [\hat X_3 - \frac{h_i'(\vartheta_i)}{\sqrt{c_i}} \hat X_i] \;\xrightarrow{P}\; 0 \;.$$

Now $\hat\vartheta^*_{S_1} \in H_0^b$ is equivalent to $h_1(\hat\vartheta^*_{S_1,1}) \ge h_2(\hat\vartheta_2)$ and hence to $\sqrt{n_3}(\hat\vartheta^*_{S_1,1} - \vartheta_1) \ge \sqrt{\frac{n_1}{c_1}}[h_1^{-1}(h_2(\hat\vartheta_2)) - h_1^{-1}(h_2(\vartheta_2))]$, where $\hat\vartheta^*_{S_1,1}$ is the first component of $\hat\vartheta^*_{S_1}$ (note that $h_1^{-1}(h_2(\vartheta_2)) = \vartheta_1$). An application of Lemma 2.5 gives

$$[h_1(\hat\vartheta^*_{S_1,1}) - h_2(\hat\vartheta_2)] \;-\; [{\Sigma_1^*}^{-1} C_1^\top \Sigma_1 \hat X_1 - \frac{(h_1^{-1}[h_2(\vartheta_2)])'}{\sqrt{c_2}} \hat X_2] \;\xrightarrow{P}\; 0 \;.$$

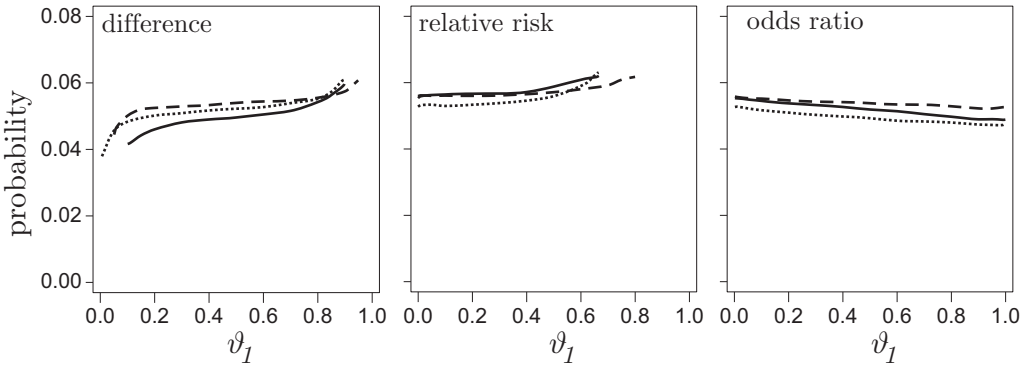

Figure 6.2: The asymptotic probability $P(T > 3.84)$ as a function of the rate ϑ_1 for several parameter constellations of $\theta_1, \theta_2, c_1, c_2$ (solid line: $0.1, 0.2, 1, 1$ for the difference, $1.5, 2, 1, 0.5$ for the relative risk and the odds ratio; dotted line: $0.1, 0.1, 1, 0.5$ for the difference, $1.5, 1.5, 1, 1$ for the relative risk and the odds ratio; dashed line: $0.05, 0.1, 0.5, 0.5$ for the difference, $1.25, 1.5, 0.5, 0.5$ for the relative risk and the odds ratio) and for hypothesis H_0^b using the difference, the relative risk and the odds ratio.

The proof for $\hat{\vartheta}^*_{S_2} \in H_0^b$ is carried out analogously.

Since $P(A \cap (B \cup C)) = P(A \cap \overline{B} \cap C) + P(A \cap \overline{C} \cap B) + P(A \cap B \cap C)$ for arbitrary events A, B, C, we find continuous functions f_{kj} and finite k, j such that

$$P(T > t) = \sum_k P\left(\bigcap_j f_{kj}(\hat{X}, t) > 0\right) + o(1) \ .$$

Theorem 2.3 gives that $\hat{X} \xrightarrow{\mathcal{D}} N_3(0, \Sigma^{-1}(\vartheta))$. Slutsky's theorem finishes the proof. □

Theorem 6.4 shows that the asymptotic distribution of the LR depends on the parameter ϑ_1 and the functions h_i for $\vartheta \in H_0^b$. The main argument used in this Theorem is the asymptotic linearity of the LR. The probability $P(T > t)$ can be calculated by simulation.

To investigate the magnitude of the dependence of the probability $P(T > t)$ in Theorem 6.4 on the nuisance parameter ϑ_1, simulations are performed for several parameter constellations. Figure 6.2 shows the asymptotic probability $P(T > 3.84)$ for the difference, the relative risk, and the odds ratio.

It is found that the asymptotic probability $P(T > 3.84)$ is stronger influenced by the nuisance parameter ϑ_1 in case of the difference than in case of the relative risk or the odds ratio, especially for small values of ϑ_1. To construct a strict level α test for all

ϑ_1, the quantile could be determined by maximizing the asymptotic probability. This approach will be not pursued for two reasons. From Theorem 6.4 it becomes apparent that the limit distribution is hardly tractable numerically, and this concept will lead to a very conservative test as can be seen from Figure 6.2.

In the following we will give a method which will be shown be numerically tractable and which gives a quite accurate approximation of the nominal level. To this end we determine the asymptotic probability for that ϑ_1 which is most likely under the null hypothesis. Therefore, the quantile t is specified such that $P_{\hat{\vartheta}_1^*}(T > t) = \alpha$, where $\hat{\vartheta}_1^*$ is the MLE of ϑ_1 constrained to $\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)$, i.e. $\hat{\vartheta}_1^*$ is the first component of $\hat{\vartheta}_{K_3}^*$ in Theorem 6.3. In Section 6.4 this approach is compared numerically to the commonly used asymptotic methods with respect to level and power.

6.3 Exact version of the LR test

The asymptotic investigations of the LR test for two samples have shown that for small sample sizes the asymptotic LR test tends to attain a somewhat liberal level. Analogously to the two-sample case, an exact version of the LR test can be constructed. This is carried out exactly as described for two samples in Section 5.3.1. In a first step, the p-values can be estimated by calculating

$$p^*(\hat{\vartheta}) = \sum_{\widetilde{\vartheta} \in \Psi} L(\hat{\vartheta}^*) \ ,$$

for any outcome $\hat{\vartheta} = (x_i/n_i)_{i=1,2,3}$, where
$\Psi := \{\widetilde{\vartheta} \in (0, \ldots, n_1) \times (0, \ldots, n_2) \times (0, \ldots, n_3) \mid T(\widetilde{\vartheta}) \geq T(\hat{\vartheta})\}$ (for T and L see (6.2) and (6.3), respectively). These are the exact p-values under the assumption that $\hat{\vartheta}^*$ (the MLE constrained to H_0^b) is the true parameter.

In a second step, these p-values $p^*(\hat{\vartheta})$ are used as an ordering criterion to define the critical region. Hence, outcomes $\hat{\vartheta}$ with corresponding small p-values are included into the critical region CR as long as

$$\max_{\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)} \sum_{\hat{\vartheta} \in CR} L(\vartheta) \leq \alpha \ .$$

In the next section this unconditional exact modification of the LR test - denoted by *exact LR test* in the following - is compared to pairwise two-sample tests.

6.4 Level and power comparisons

The LR principle for the hypotheses H_0^a and H_0^c leads to a combination of the two-sample LR tests, where no level adjustment is necessary. Therefore, no further investigations are carried out for these hypotheses and we refer to the investigations for the two-sample case in Section 5.3.3. However, as seen in Section 6.2, for the hypothesis H_0^b the LR test cannot be reduced to the two-sample case. This will be investigated more precisely in the following.

The exact LR test is compared with the two-sample LR tests and with Chan's unconditional exact two-sample tests. The pairwise two-sample tests are level adjusted applying Bonferroni's and Hochberg's procedure, respectively. Even if Hochberg's approach does not guarantee a strict level α test (dependency of the pairwise test statistics), we included it into the comparison, since we found numerically quite satisfactory results. The level adjustment from Dunnett (cf. Section 4.2) is omitted, since for Dunnett's approach normally distributed data have to be assumed.

The exact procedures are investigated for small sample sizes (up to 50 per group). Note that the computation time (approximately 20 minutes for a sample size of 25 per group and 120 minutes for 50 per group with a Pentium III, 1.2 MHz, SAS V8) increases rapidly for larger sample sizes. For sample sizes between 50 and 500 per group the asymptotic LR test is numerically compared with the commonly used asymptotic two-sample tests (see Section 5.2.2) applying Bonferroni's and Hochberg's adjustment.

The power is exactly calculated as in Section 5.3.3 comparing the exact LR test and its competitors. A broad scenario of parameter settings $(\theta_1, \theta_2, n_1, n_2, n_3, \vartheta_1)$ is considered for the distance measures difference, relative risk and odds ratio:

- *Equivalence margins*:
  $(\theta_1, \theta_2) \in \{(0.15, 0.15), (0.15, 0.2), (0.15, 0.25), (0.2, 0.2), (0.2, 0.25), (0.25, 0.25)\}$
  for h_{DI} and $(\theta_1, \theta_2) \in \{(1.5, 1.5), (1.5, 2), (1.5, 2.5), (2, 2), (2, 2.5), (2.5, 2.5)\}$ for h_{RR} and h_{OR} .
- *Sample size*: Balanced sample sizes
  $n_1 = n_2 = n_3 \in \{20, 25, 30, 40, 50\}$ and unbalanced sample sizes
  $(n_1, n_2, n_3) \in \{(20, 20, 40), (25, 25, 50), (40, 40, 20), (50, 50, 25)\}$.
- *Nuisance parameter*: $\vartheta_1 \in \{0.1, 0.2, 0.3, 0.5, 0.8, 0.9\}$.
- *Distances between the groups*: $\vartheta_1 = \vartheta_2 \leq \vartheta_3$ and $\vartheta_1 \geq \vartheta_2 = \vartheta_3$.

Overall, 648 parameter constellations for each distance measure, respectively, are considered in a first step. Configurations are omitted in case of non-feasible settings (e.g.

$\vartheta_1 \geq 1-\theta_1$ for h_{DI}, $\vartheta_1 \geq 1/\theta_1$ for h_{RR}). The parameters $\vartheta_1, \vartheta_2, \vartheta_3$ are chosen such that, if possible, the resulting power is larger than 0.8, at least for one of the tests compared. Here two settings are investigated: either the parameter ϑ_3 is chosen equal to or smaller than $\vartheta_1 = \vartheta_2$, or the parameter ϑ_1 is chosen equal to or greater than $\vartheta_2 = \vartheta_3$. In a second step, parameter constellations are omitted for which all tests achieve a power larger than 0.9. Finally, 261 parameter constellations remain for the difference, 85 parameter constellations remain for the relative risk, and 260 parameter constellations remain for the odds ratio.

Figures 6.3, 6.4 and 6.5 represent the power of the exact LR test (vertical axes) and its competitors (horizontal axes) for the three distance measures h_{DI}, h_{RR} and h_{OR}, respectively. The calculations show that the power of the exact LR test compared to its pairwise competitors is larger in general. For nearly each parameter constellation and distance measure the LR test outperforms the pairwise procedures using Bonferroni's adjustment. In comparison to Hochberg's adjustment the power improvement is smaller. Figure 6.6 gives Boxplots (results of all distance measures combined) of the power differences between the exact LR test and its competitors for Bonferroni's and Hochberg's adjustment, respectively. The power enhancement using the exact LR test is rather substantial in comparison to Bonferroni's adjustment.

The power of the LR test also tends to be larger (similar to the results of the two-sample case) compared to Hochberg's adjustment. However, the improvement is surprisingly low.

We found no parameter constellation for which Hochberg's adjustment led to an increased level larger than α (data not displayed). However, recall that there is no rigid result which reveals Hochberg's test as a level α test. The reason for the good performance of the pairwise two-sample procedures using Hochberg's adjustment is illustrated in Figure 6.7. Here the differences between the rejection regions of the exact LR test and Chan's test are displayed for an arbitrary setting. The small dots represent the outcomes x_1, x_2, x_3 for which the exact 3-sample LR test leads to the rejection of H_0^b only, the big dots represent the outcomes for which Chan's test leads to the rejection of H_0^b only. On the left hand side Chan's test with Hochberg's adjustment is applied, on the right hand side Bonferroni's adjustment is applied. It is found that considerably more outcomes are included into the critical region applying the LR test. However, those outcomes which have largest probability under alternatives of interest (e.g. when $\vartheta_1 = \vartheta_2 = \vartheta_3$) are placed at the "edge" ($x_1 \approx x_2 \approx x_3$). Hence, outcomes which are not placed near the "edge" will provide a minor contribution to the power.

As a conclusion, the calculations show that the exact LR test improves the power compared to the pairwise two-sample procedures. The improvement is quite substantial compared to the Bonferroni adjusted procedures. Compared to the Hochberg adjusted proce-

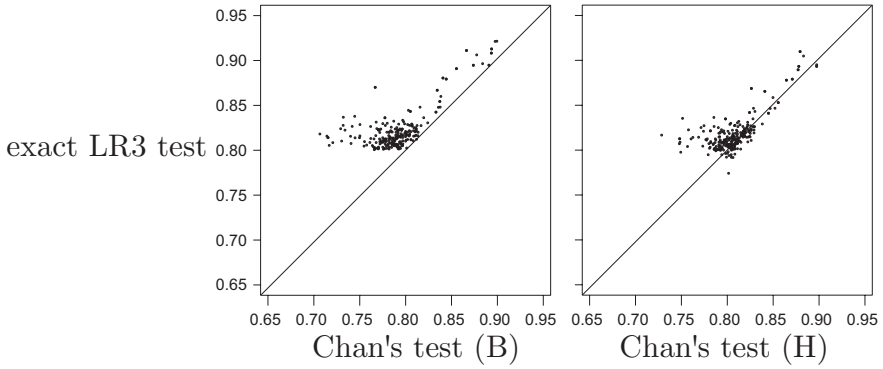

Figure 6.3: The power of the exact 3-sample LR test (vertical axis) in comparison to the pairwise 2-sample tests by Chan with Bonferroni's (B) and Hochberg's (H) adjustment (horizontal axis) for several parameter constellations and for hypothesis H_0^b using the difference.

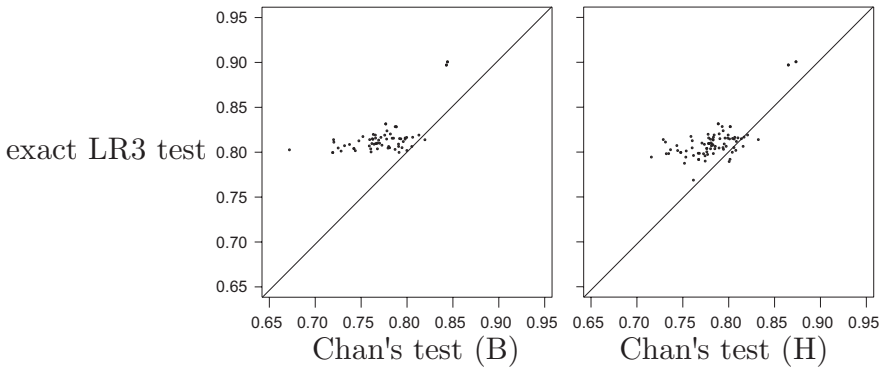

Figure 6.4: The power of the exact 3-sample LR test (vertical axis) in comparison to the pairwise 2-sample tests by Chan with Bonferroni's (B) and Hochberg's (H) adjustment (horizontal axis) for several parameter constellations and for hypothesis H_0^b using the relative risk.

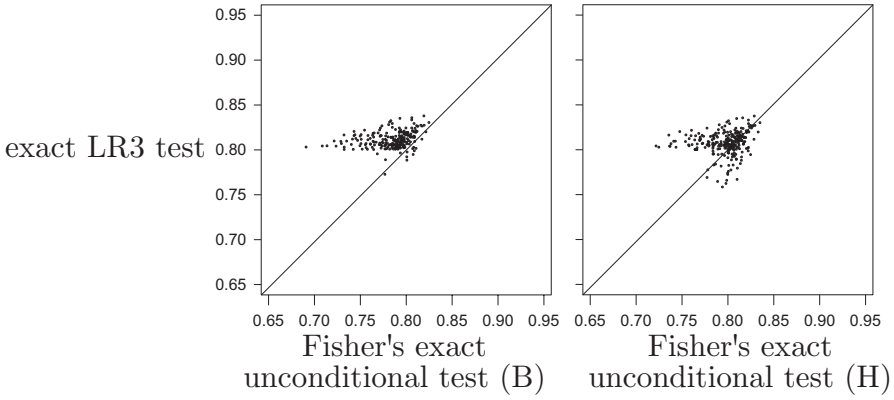

Figure 6.5: The power of the exact 3-sample LR test (vertical axis) in comparison to the pairwise 2-sample unconditional exact tests by Fisher with Bonferroni's (B) and Hochberg's (H) adjustment (horizontal axis) for several parameter constellations and for hypothesis H_0^b using the odds ratio.

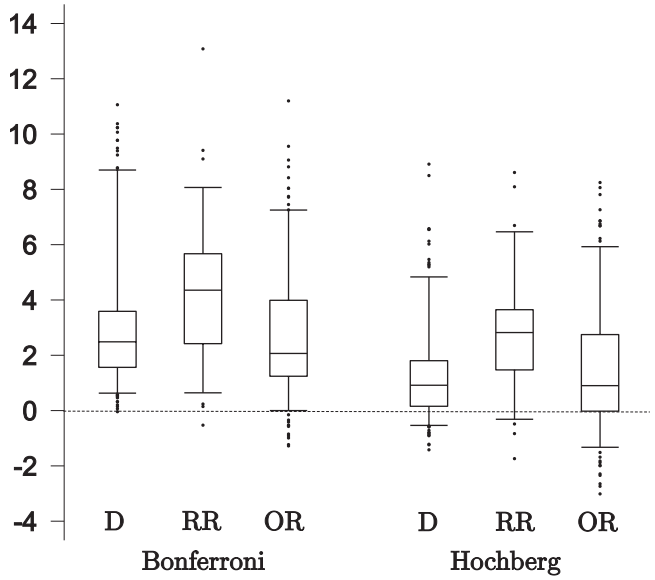

Figure 6.6: Boxplot (whiskers are the 5% and 95% quantiles) for the power differences (times 100) between the exact LR test and the multiple comparison procedures using Bonferroni's and Hochberg's adjustment for the three distance measures difference (D), relative risk (RR) and odds ratio (OR).

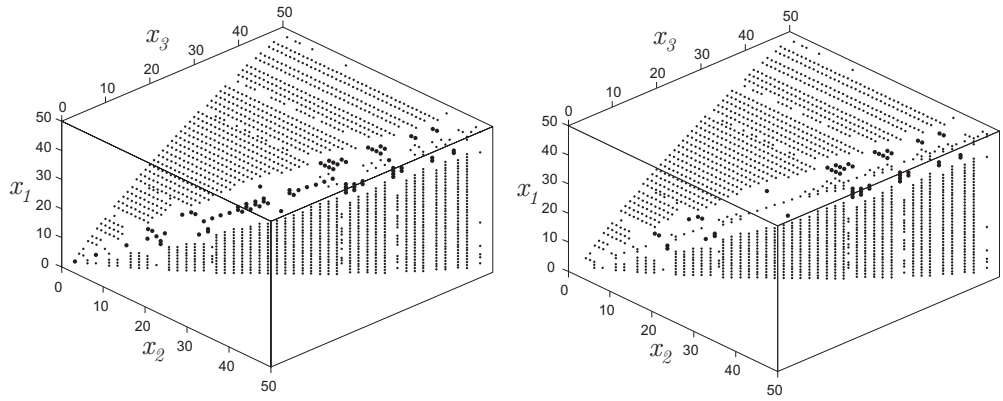

Figure 6.7: The differences between the rejection regions of the exact LR test and Chan' test with Hochberg's adjustment (left hand side) and Bonferroni's adjustment (right hand side). The small dots represent the outcomes x_1, x_2, x_3 for which the exact 3-sample LR test leads to the rejection of H_0^b only, the big dots represent the outcomes for which Chan's test leads to the rejection of H_0^b only ($n_1 = n_2 = n_3 = 50$, $\theta_1 = \theta_2 = 0.15$).

dures the improvement tends to be slightly smaller, but it should be taken into account that Hochberg's adjustment does not guarantee to keep the level α in case of hypothesis H_0^b.

The sample size can be significantly reduced applying the exact LR test instead of the Bonferroni adjusted procedure. This will be illustrated by the following example. Let $(\theta_1, \theta_2) = (2, 2)$ and $(\vartheta_1, \vartheta_2, \vartheta_3) = (0.8, 0.8, 0.6)$. Then a sample size of 25 per group is required to give a power of 0.8 when using the Bonferroni adjusted procedure. Applying the exact LR test, a sample size of 20 per group yields a power of 0.8, i.e. the sample size can be reduced by 20%. As mentioned in the two-sample case, the sample size may even be further reduced choosing unequal group sample sizes.

The asymptotic LR test is numerically investigated by simulations (100,000 repetitions) for sample sizes between 50 and 500 per group. Level and power of the LR test is compared to the pairwise asymptotic two-sample tests based on the score statistic which are introduced in Section 5.2.2. The pairwise tests are applied using Bonferroni's and Hochberg's adjustment. We have seen for normally distributed data that Hochberg's adjustment keeps the level α. Therefore, it is interesting to investigate this for binomial data.

Table 6.1 shows the simulated level and power for the three distance measures difference, relative risk and odds ratio, respectively. Different parameter settings are implemented, analogously to the investigations mentioned above. The simulated levels are quite accu-

rate for all approaches. Overall, it can be seen that the LR test is almost always slightly superior to its competitors with respect to level and power.

6.5 Example

In a randomized double-blind comparison in patients with cancer Hesketh et al. [1996] assess the efficacy of antiemetic agents in preventing cisplatin-induced nausea and vomiting. The trial was performed to show non-inferiority of dolasetron mesylate at doses of 1.8 mg/kg (T_1) and 2.4 mg/kg (T_2), respectively, over the standard ondansetron (C) at its approved dose of 32 mg. The primary analysis was done by comparing the failure rates of T_1 and T_2, respectively, with C. Patients having emetic episodes or receiving rescue medication during 24 hours were classified as failures. For both comparisons the equivalence margin for the odds ratio was specified as 2. It is not clearly described by Hesketh et al. [1996] whether it was the goal to show non-inferiority of both doses of dolasetron compared to ondansetron, or to show that at least one of the doses of dolasetron is non-inferior to ondansetron.

The resulting failure rates were similar in the three groups: 110/198 (56%) in T_1, 123/205 (60%) in T_2, and 118/206 (57%) in C. Comparing T_1 versus C and T_2 versus C, the authors calculated an odds ratio (upper 97.5% confidence limit) of 0.97 (1.47) and 1.16 (1.75), respectively. They concluded that dolasetron (1.8 or 2.4 mg/kg) has comparable efficacy to ondansetron, since the upper confidence limits were smaller than 2 (without specifying any level adjustment).

If we apply the asymptotic LR test (which equals the pairwise comparisons with level α, respectively) for H_0^a, i.e. for showing that at least one of the treatments T_1, T_2 is non-inferior to C, we obtain p-values of 0.00007 and 0.0019 comparing T_1 versus C and T_2 versus C, respectively. The same p-values result for the asymptotic score tests (see Section 5.2.2). We can determine test-based upper 97.5% confidence limits by calculating the hypotheses boundaries for which the LR tests do not reject the null hypotheses at level 2.5%. This results in boundaries 1.38 for T_1 versus C and 1.66 for T_2 versus C which are even a bit smaller than the boundaries given by Hesketh et al. [1996]. Even if their boundaries - calculated with adjustment for covariates - are not directly comparable to our boundaries, this indicates how powerful the LR test is. If it is of interest to show non-inferiority of both doses of dolasetron compared to ondansetron, we can apply the asymptotic LR test for H_0^b. For this example we get $T = 15.9$ (the test statistic (6.3)), which is larger than the simulated quantile $t = 3.81$. The approximated p-value is smaller than 0.0001.

Table 6.1: The simulated power (level) (times 100) of the asymptotic LR test and its corresponding asymptotic pairwise score tests using Bonferroni's (B) and Hochberg's (H) adjustment for different parameter constellations and distance measures ($\tilde{\theta}_1, \tilde{\theta}_2$ as the true differences, relative risks, or odds ratios, respectively).

n_1	n_2	n_3	θ_1	θ_2	ϑ_1	$\tilde{\theta}_1$=$\tilde{\theta}_2$	LR test	score test (B)	score test (H)
						difference			
50	50	100	0.1	0.15	0.2	-0.02	84.9 (5.0)	84.7 (5.0)	85.2 (5.1)
75	75	75	0.1	0.15	0.3	-0.05	85.3 (5.0)	83.6 (4.7)	84.8 (4.8)
75	75	150	0.1	0.15	0.25	0	80.5 (4.9)	79.2 (5.0)	80.7 (5.2)
100	100	100	0.1	0.15	0.2	0	80.9 (5.1)	77.8 (4.2)	79.5 (4.7)
100	100	200	0.1	0.15	0.3	0	85.7 (4.9)	84.5 (4.8)	85.6 (5.1)
200	200	200	0.1	0.1	0.25	0	80.7 (5.0)	78.1 (4.3)	79.6 (4.7)
200	200	400	0.1	0.1	0.4	0	86.4 (4.9)	84.3 (4.8)	85.2 (5.0)
400	400	400	0.05	0.1	0.5	0	83.8 (4.9)	82.9 (4.7)	83.3 (4.6)
500	500	500	0.05	0.1	0.5	0	90.6 (5.3)	90.0 (4.7)	90.5 (5.2)
						relative risk			
50	50	100	1.5	1.75	0.4	1	82.4 (4.8)	79.8 (5.2)	81.2 (5.2)
75	75	75	1.5	1.75	0.35	1	80.9 (4.9)	79.1 (5.2)	80.2 (5.1)
75	75	150	1.5	1.75	0.3	1	81.8 (5.0)	80.6 (5.2)	81.6 (5.5)
100	100	100	1.5	1.5	0.2	0.75	81.5 (5.0)	80.3 (4.9)	81.6 (5.2)
100	100	200	1.5	1.5	0.3	1	79.4 (5.0)	77.1 (5.2)	78.6 (5.2)
200	200	200	1.25	1.5	0.3	1	80.0 (5.1)	78.9 (4.9)	80.0 (5.2)
200	200	400	1.25	1.5	0.25	1	80.7 (4.8)	79.8 (5.1)	80.5 (5.1)
400	400	400	1.25	1.25	0.35	1	83.6 (4.9)	80.7 (4.6)	82.3 (4.8)
500	500	500	1.25	1.25	0.3	1	82.8 (4.8)	80.7 (4.6)	81.8 (4.7)
						odds ratio			
50	50	100	1.5	1.75	0.4	0.75	80.2 (4.7)	77.6 (4.9)	78.9 (4.9)
75	75	75	1.5	1.75	0.5	0.8	77.8 (4.9)	75.3 (4.5)	76.9 (4.7)
75	75	150	1.5	1.75	0.3	0.8	83.0 (4.9)	80.6 (5.0)	81.8 (5.2)
100	100	100	1.5	1.5	0.4	0.75	83.9 (5.3)	82.2 (4.8)	83.7 (5.2)
100	100	200	1.5	1.5	0.45	0.85	84.5 (4.9)	82.4 (4.9)	83.3 (5.2)
200	200	200	1.25	1.5	0.3	0.85	79.8 (5.0)	78.2 (4.6)	79.1 (4.9)
200	200	400	1.25	1.5	0.25	0.9	79.5 (5.0)	77.6 (4.9)	78.8 (5.2)
400	400	400	1.25	1.25	0.35	0.9	79.0 (5.1)	75.5 (4.6)	77.4 (4.9)
500	500	500	1.25	1.25	0.2	0.85	84.3 (4.9)	81.8 (4.5)	83.1 (4.7)

7 Conclusions

In this work the LR test is investigated for hypotheses which allow to decide whether a treatment group is non-inferior or relevantly superior to a control group. We have assumed that the data follow a normal or binomial distribution.

In the first part the LR test is derived for normally distributed data comparing the means. For the two-sample case we have shown that the LR test is equivalent to the well known t-test when using the difference or the ratio of the means. For three-group comparisons we have investigated various hypotheses which are of interest in medical research. As seen, some of these hypotheses lead to the intersection-union test when applying the LR principle. For other hypotheses the LR test results in tests well known from order restricted inference. Finally, we have compared this method to commonly used multiple comparison procedures (pairwise comparisons using Bonferroni's, Dunnett's, and Hochberg's level adjustment). The simulation results show that the LR test is similar (but slightly superior) to the pairwise procedures with respect to power.

The main contribution of this work concerns the comparison of two and three binomial distributions. The asymptotic methodology is derived for the LR test with respect to smooth boundary functions of the hypotheses. In the two-sample case the asymptotics for the likelihood ratio test for general hypotheses is found to follow a $\frac{1}{2} + \frac{1}{2}\chi^2_1$-law. We have seen that for small sample sizes this is not reliable. Nevertheless, for sample sizes larger than 100, say, the asymptotic test yields quite accurate results. Our comparison of the LR test to its asymptotic competitors based on score statistics shows that the power is similar, irrespective of the chosen distance measure. However, we found that the LR test keeps its nominal level more accurately than the competitors.

For small sample sizes we have derived an unconditional exact approach using the LR statistic (exact LR test), and we have compared it to other unconditional exact methods mentioned in the literature. Since a comprehensive comparison of unconditional exact tests was never given in the literature, an extensive power investigation is performed. For a broad scenario of parameter settings the exact power of the exact LR test is compared to the commonly used unconditional exact approaches: Chan's test, the π_{local} test, and an unconditional version of Fisher's exact test. Irrespective of the chosen distance measure, the power of the exact LR test tends to be larger, even if the improvement is small

in general. As a by-product we found that Barnard's test leads to intrinsic numerical difficulties. Therefore, this test cannot be proposed in practice.

In the last part of this work the asymptotic distribution of the LR test for three binomial distributions and general boundary functions of the hypotheses is derived. As for normally distributed data, the LR principle leads to multiple pairwise comparisons for some of the hypotheses. However, for other hypotheses we obtain a test which is different from multiple comparison procedures. We have shown that in this case the asymptotic distribution is rather complicated. In particular, it depends on an unknown parameter and cannot be used in practise. Therefore, we have proposed to estimate this parameter under constraints which gives satisfactory results.

Finally, for small sample sizes we have analyzed an exact version based on the LR statistic. A broad comparison to pairwise exact procedures shows that the power can be substantially improved applying the exact LR test.

We have briefly discussed sample size calculations for the two-sample case. It is shown that the power is not maximized by balanced sample sizes in case of non-inferiority trials. An unequal allocation of patients may lead to a substantial power improvement. It would be a challenging task for further research to investigate this topic more comprehensively, and to extend it to more than two groups. Another task which is only briefly addressed in this work will be the derivation of confidence intervals based on the statistical tests investigated here. Especially the comparison of confidence intervals based on the exact procedures for shifted hypotheses to the commonly applied approaches would be of interest. Here shorter confidence intervals are to be expected.

A Symbols and abbreviations

Symbol	Explanation (page of first occurrence)
$\mathbb{R}$	real numbers (11)
$\mathbb{N}$	integers (12)
$f'(x)$, $f''(x)$	first, second derivative (13)
$A^\top$	transposed matrix (13)
$diag(x_1, \ldots, x_k)$	diagonal matrix (62)
$(x_i)_{i=1,\ldots,k}$	vector $(x_1, \ldots, x_k)^\top$ (58)
$\|\|x\|\|$	Euclidian norm of a vector (12)
$\partial\Theta$	boundary of a set Θ (12)
$F_{\chi^2_1}$	cumulative distribution function of the Chi-squared distribution with 1 degree of freedom (37)
$(F)_\alpha$	α-quantile of a distribution F (15)
$\overset{i.i.d.}{\sim}$	independent identically distributed (15)
$B_{n,m}$	Beta-distribution (26)
$Bi(n, \vartheta)$	Binomial distribution (33)
χ^2_k	Chi-squared distribution with k degrees of freedom (14)
$N(\mu, \sigma^2)$, $N_c(\mu, \Sigma)$	univariate, c-variate normal distribution with mean μ and variance σ^2 or covariance matrix Σ (13)
u_α	lower α-quantile of the standard normal distribution (18)
t_m, $t_{m,\delta}$	central, noncentral t-distribution with noncentrality parameter δ and m degrees of freedom (15)
$\hat{\vartheta}$	unconditional ML estimator (11)
$\hat{\vartheta}^*$	conditional ML estimator (11)
$\xrightarrow{\mathcal{D}}$	convergence in distribution (13)
$\xrightarrow{P}$	convergence in probability (13)

Abbreviations:

CR	Critical region of a statistical test
IUT	Intersection-union test
LR	Likelihood ratio
MLE	Maximum likelihood estimator

B Tables

Table B.1: Difference: The exact power γ (times 100) for parameter constellations which gives the most extreme power differences between the exact LR test and the best of its competitors ($|\gamma_{exact\,LR} - \max\{\gamma_{competitors}\}| > 0.015$) .

θ_0	n_1	n_2	ϑ_2	ϑ_1	exact LR	Chan	π_{local}
0.15	35	35	0.1	0.07	81.1	77	71.3
0.05	100	60	0.9	0.8	81.3	77.3	77.6
0.1	50	50	0.1	0.06	80.2	77	75.8
0.1	60	60	0.1	0.06	85.8	82.7	81.8
0.05	60	30	0.9	0.73	80.3	75.2	77.5
0.15	50	50	0.1	0.09	82.5	80	76.9
0.05	50	25	0.9	0.7	82	79.5	79.5
0.2	30	20	0.1	0.08	84.5	82.1	72.9
0.05	25	25	0.9	0.66	80.7	78.4	77.8
0.05	35	35	0.3	0.11	81	79	79
0.2	25	25	0.2	0.12	83.2	81.2	80.6
0.2	25	25	0.3	0.19	80.9	78.9	78.9
0.25	50	50	0.3	0.3	86.1	84.1	84.1
0.05	35	35	0.9	0.71	82.4	80.5	80.5
0.05	60	30	0.1	0.01	88.4	86.5	86.1
0.15	100	100	0.2	0.2	83.9	82	81.2
0.15	100	100	0.8	0.8	83.9	82	81.2
0.1	100	100	0.1	0.09	82.4	80.6	78.3
0.15	50	50	0.2	0.15	83	81.3	79.9
0.15	35	35	0.8	0.68	81.3	79.7	79.7
0.15	35	35	0.3	0.18	83.2	81.6	81.5
0.1	60	60	0.2	0.12	84.3	82.6	82.7
0.05	60	30	0.2	0.06	82.6	84.3	82.6
0.05	80	60	0.1	0.04	79.7	81.6	78.9
0.2	30	20	0.2	0.13	79.1	81.8	80.4
0.2	30	20	0.3	0.19	79.4	82.4	82.5
0.05	50	50	0.1	0.02	82.5	85.7	82.3
0.1	40	40	0.1	0.04	81.5	85.3	81.5

Table B.2: Relative risk: The exact power γ (times 100) for parameter constellations which gives the most extreme power differences between the exact LR test and the best of its competitors ($|\gamma_{exact\,LR} - \max\{\gamma_{competitors}\}| > 0.015$) .

θ_0	n_1	n_2	ϑ_2	ϑ_1	exact LR	Chan	π_{local}
1.1	60	30	0.3	0.09	84	78.4	81.2
1.5	60	40	0.2	0.07	82.8	78.1	80
2.5	100	50	0.1	0.04	82.3	74.1	79.5
2.5	50	25	0.2	0.09	81.1	75.3	78.4
1.1	100	50	0.9	0.81	84.1	74.9	81.8
1.1	80	40	0.3	0.12	81.6	76.1	79.4
1.1	60	60	0.8	0.64	87.6	85.7	85.7
2.5	30	30	0.3	0.24	81.4	79.5	79.6
2.5	80	40	0.1	0.025	82.4	76.6	80.6
1.1	100	100	0.9	0.855	84.7	82.3	83
1.1	100	60	0.9	0.81	88.7	86.9	86.9
1.1	60	40	0.9	0.765	86.5	82.2	84.8
1.1	60	30	0.9	0.72	88.5	81.3	86.8
1.1	30	20	0.9	0.675	80	74.3	78.4
1.25	60	60	0.5	0.35	85.5	83.2	84
1.5	30	30	0.3	0.12	78.5	78.5	80.2
1.25	25	25	0.5	0.225	82.1	83.9	83.9
1.1	80	80	0.1	0.015	80.4	82.3	78.4
2	30	30	0.3	0.18	79.2	78.5	81.3
2	80	60	0.1	0.035	78.4	81.3	78.3
2	25	25	0.2	0.04	78.7	82.2	76.2
1.5	40	40	0.2	0.05	81.2	84.7	81.1
2	40	20	0.2	0.04	77.6	81.7	77.6
2	80	40	0.1	0.02	76.1	80.4	75.8
2.5	30	20	0.2	0.05	78.4	82.8	78.4
1.1	80	60	0.1	0.01	78.6	83.4	78.6
1.5	60	60	0.1	0.015	76.7	81.6	76.7

Table B.3: Odds ratio: The exact power γ (times 100) for parameter constellations which gives the most extreme power differences between the exact LR test and the best of its competitors ($|\gamma_{exact\,LR} - \max\{\gamma_{competitors}\}| > 0.03$) .

θ_0	n_1	n_2	ϑ_2	ϑ_1	exact LR	Fisher's exact unconditional	π_{local}
1.25	100	50	0.1	0.011	85.4	77.3	77.3
2.5	40	40	0.1	0.016	81.6	74.3	74.3
1.5	80	50	0.1	0.016	82.2	75.1	75.1
2.5	50	50	0.1	0.027	81.7	74.7	74.7
1.1	35	35	0.2	0.024	83.8	77.2	77.2
2.5	20	20	0.2	0.036	81.8	75.2	75.2
1.25	30	30	0.2	0.024	80.4	74.3	74.3
2	80	80	0.1	0.043	80	74.7	74.7
1.1	60	60	0.2	0.059	81.2	76.4	76.4
2.5	80	40	0.1	0.022	82.6	78.1	78.1
2	50	50	0.1	0.011	85.2	80.7	80.7
1.5	100	60	0.1	0.022	83.2	78.9	78.9
1.25	35	35	0.2	0.024	86.3	82	82
1.1	60	30	0.2	0.024	86.6	82.5	82.5
2.5	100	60	0.1	0.048	80.8	76.8	76.8
2	80	50	0.1	0.027	81.1	77.2	76.4
2	25	25	0.2	0.024	84.9	81.1	81.1
1.1	100	80	0.1	0.016	83.8	80.2	80.2
1.25	40	40	0.2	0.036	84.4	80.8	80.8
2	100	100	0.1	0.053	81.2	77.8	77.8
1.1	40	40	0.2	0.036	80.5	77.1	77.1
1.25	80	60	0.1	0.011	85.5	82.1	82.1
1.25	60	60	0.2	0.07	80.3	77	77
2.5	60	60	0.1	0.037	80.7	77.4	77.4
1.1	50	50	0.2	0.048	81.5	78.3	78.3
2.5	100	80	0.1	0.058	82.5	79.3	78.8
1.25	100	60	0.1	0.016	80.2	77.1	77.1
2	80	50	0.2	0.121	80.5	77.5	77.3
1.5	35	35	0.2	0.036	81.6	84.6	84.6

Bibliography

ASSENT (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: The assent-2 double-blind randomised trial. Assessment of the safety and efficacy of a new thrombolytic investigators. *The Lancet*, 354:716–722.

Barlow R. E., Bartholomew D. J., Bremner J. M., and Brunk H. D. (1972). *Statistical Inference under Order Restrictions*. John Wiley & Sons, London.

Barnard G. A. (1945). A new test for 2x2 tables. *Nature*, 156:177.

Barnard G. A. (1947). Significance tests for 2x2 tables. *Biometrika*, 34:123–138.

Berger R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300.

Blackwelder W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials*, 3:345–353.

Blackwelder W. C. (1993). Sample size and power for prospective analysis of relative risk. *Statistics in Medicine*, 12:691–698.

Boschloo R. D. (1970). Raised conditional level of significance for the 2x2 table when testing the equality of two probabilities. *Statistica Neerlandica*, 24:1–35.

Bristol D. R. (1996). Determining equivalence and the impact of sample size in anti-infective studies: A point to consider. *Journal of Biopharmaceutical Statistics*, 6: 319–326.

Casella G. and Berger R. L. (2002). *Statistical Inference*. Duxbury, USA, 2nd edition.

Chan I. S. F. (1998). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine*, 17:1403–1413.

Chen J. J., Tsong Y., and Kang S.-H. (2000). Tests for equivalence or noninferiority between two proportions. *Drug Information Journal*, 34:569–578.

Childs D. R. (1967). Reduction of the multivariate normal integral to characteristic form. *Biometrika*, 54:293–300.

Chouela E. N., Abeldano A. M., Pellerano G., Forgia M. La, Papale R. M., Garsd A., Balian M. C., Battista V., and Poggio N. (1999). Equivalent therapeutic efficacy and safety of ivermectin and lindane in the treatment of human scabies. *Archives of Dermatology*, 135:651–655.

Chuang-Stein C. (2001). Testing for superiority or inferiority after concluding equivalence? *Drug Information Journal*, 35:141–143.

Committee for Proprietary Medicinal Products (1997). Note for guidance on evaluation of new anti-bacterial medicinal products.

Committee for Proprietary Medicinal Products (1999). Note for guidance on clinical evaluation of new vaccines.

D'Agostino R. B., Chase W., and Belanger A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician*, 42:198–202.

D'Agostino R. B. and Heeren T. C. (1991). Multiple comparisons in over-the-counter drug clinical trials with both positive and placebo controls. *Statistics in Medicine*, 10: 1–6.

Dammann H. G., Folsch U. R., Hahn E. G., Kleist von D. H., Klor H. U., Kirchner T., Strobel S., and Kist M. (2000). Eradication of h. pylori with pantoprazole, clarithromycin, and metronidazole in duodenal ulcer patients: A head-to-head comparison between two regimens of different duration. *Helicobacter*, 5:41–51.

Diehm C., Trampisch H. J., Lange S., and Schmidt C. (1996). Comparison of leg compression stocking and oral horse-chestnut seed extract therapy in patients with chronic venous insufficiency. *Lancet*, 347:292–294.

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

Dunnett C. W. and Gent M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics*, 33: 593–602.

Dunnett C. W. and Tamhane A. C. (1997). Multiple testing to establish superiority/equivalence of a new treatment compared with k standard treatments. *Statistics in Medicine*, 16:2489–2506.

Farrington C. P. and Manning G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9:1447–1454.

FDA (1992). Points to consider. Clinical development and labeling of anti-infective drug products.

FDA (1998). Guidance for industry. Complicated urinary tract infections and pyelonephritis - developing antimicrobial drugs. Draft guidance.

Ferguson T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall.

Finner H. and Strassburger K. (2001). Ump(u)-tests for a binomial parameter: A paradox. *Biometrical Journal*, 43:667–675.

Gart J. J. (1971). The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Review of the International Statistical Institute*, 39:148–169.

Greco D., Salmaso S., Mastrantonio P., Giuliano M., Tozzi A. E., Anemona A., Atti degli M. L. Ciofi, Giammanco A., Panei P., Blackwelder W. C., Klein D. L., and Wassilak S. G. (1996). A controlled trial of two acellular vaccines and one whole-cell vaccine against pertussis. Progetto Pertosse Working Group. *New England Journal of Medicine*, 334:341–348.

Gustafsson L., Hallander H. O., Olin P., Reizenstein E., and Storsaeter J. (1996). A controlled trial of a two-component acellular, a five-component acellular, and a whole-cell pertussis vaccine. *New England Journal of Medicine*, 334:349–355.

GUSTO (1993). An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. The GUSTO investigators. *New England Journal of Medicine*, 329:673–682.

Hauschke D., Kieser M., Diletti E., and Burke M. (1999). Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Statistics in Medicine*, 18:93–105.

Hesketh P., Navari R., Grote T., Gralla R., Hainsworth J., Kris M., Anthony L., Khojasteh A., Tapazoglou E., Benedict C., and Hahne W. (1996). Double-blind, randomized comparison of the antiemetic efficacy of intravenous dolasetron mesylate and intravenous ondansetron in the prevention of acute cisplatin-induced emesis in patients with cancer. dolasetron comparative chemotherapy-induced emesis prevention group. *Journal of Clinical Oncology*, 14:2242–2249.

Hochberg Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802.

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

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

Hoover D. R. and Blackwelder W. C. (2001). Allocation of subjects to test null relative risks smaller than one. *Statistics in Medicine*, 20:3071–3082.

Hsu J. C. (1996). *Multiple comparisons - Theory and methods*. Chapman and Hall, London.

ILAE (1998). Considerations on designing clinical trials to evaluate the place of new antiepileptic drugs in the treatment of newly diagnosed and chronic patients with epilepsy. Report of the ilae commission on antiepileptic drugs. *Epilepsia*, 39:799–803.

InTIME-II (2000). Intravenous npa for the treatment of infarcting myocardium early; InTIME-II, a double-blind comparison of single-bolus lanoteplase vs accelerated alteplase for the treatment of patients with acute myocardial infarction. *European Heart Journal*, 21:2005–2013.

Johnson N. L. and Welch B. L. (1940). Applications of the non-central t-distribution. *Biometrika*, 31:362–389.

Kang S. H. and Chen J. J. (2000). An approximate unconditional test of non-inferiority between two proportions. *Statistics in Medicine*, 19:2089–2100.

Kieser M. (1995). A confirmatory strategy for therapeutic equivalence trials. *International Journal of Clinical Pharmacology and Therapeutics*, 33:388–390.

Lehmann E. L. (1986). *Testing Statistical Hypotheses*. Springer-Verlag, New York, 2nd edition.

Liu J. P. and Weng C. S. (1994). Evaluation of log-transformation ni assessing bioequivalence. *Communications in Statistics - Theory and Methods*, 23:421–434.

Martín Andrés A. and Herranz Tejedor I. (2003). Exact unconditional non-classical tests on the difference of two proportions. *Computational Statistics and Data Analysis*, (to appear).

Martín Andrés A. and Silva Mato A. (1994). Choosing the optimal unconditioned test for comparing two independent proportions. *Computational Statistics and Data Analysis*, 17:555–574.

McDonald L. L., Davis B. M., and Milliken G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2x2 contingency tables. *Technometrics*, 19:145–157.

Miettinen O. and Nurminen M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, 4:213–226.

Moliterno D. J. and Topol E. J. (2000). A direct comparison of tirofiban and abciximab during percutaneous coronary revascularization and stent placement: Rationale and design of the target study. *American Heart Journal*, 140:722–726.

Mood A. M., Graybill F. A., and Boes D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, Singapore, third edition.

Moulton L. H., O'Brien K. L., Kohberger R., Chang I., Reid R., Weatherholtz R., Hackell J. G., Siber G. R., and Santosham M. (2001). Design of a group-randomized streptococcus pneumoniae vaccine trial. *Controlled Clinical Trials*, 22:438–452.

Newcombe R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, 17:873–890.

Phillips K. F. (2003). A new test of non-inferiority for anti-infective trials. *Statistics in Medicine*, 22:210–212.

Pigeot I., Schäfer J., Röhmel J., and Hauschke D. (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*, 22:883–899.

Pruscha H. (2000). *Vorlesungen über Mathematische Statistik*. B. G. Teubner, Stuttgart.

Röhmel J. and Mansmann U. (1999a). Letter to the Editor: Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies by I. S. F. Chan, Statistics in Medicine, 17, 1403-1413 (1998). *Statistics in Medicine*, 18:1734–1737.

Röhmel J. and Mansmann U. (1999b). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal*, 41:149–170.

Robertson T. and Wright F. T. (1981). Likelihood ratio tests for and against a stochastic ordering between multinomial populations. *The Annals of Statistics*, 9:1248–1257.

Robertson T., Wright F. T., and Dykstra R. L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester.

Rodary C., Com-Nougue C., and Tournade M. F. (1989). How to establish equivalence between treatments: A one-sided clinical trial in paediatric oncology. *Statistics in Medicine*, 8:593–598.

Roebruck P. and Kühn A. (1995). Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine*, 14:1583–1594.

Sasabuchi S. (1980). A test of a multivariate normal mean with composite hypotheses determined by linear inequalities. *Biometrika*, 67:429–439.

Skipka G. and Trampisch HJ. Unconditional exact tests for comparing two independent proportions. In Kunert J and Trenkler G, editors, *Festschrift in Honour of Siegfried Schach: Mathematical Statistics with Applications in Biometry*, pages 189–196. Eul publishers (2001).

Storer B. E. and Kim C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association*, 85:146–155.

Tebbe U., Michels R., Adgey J., Boland J., Caspi A., Charbonnier B., Windeler J., Barth H., Groves R., Hopkins G. R., Fenell W., Betriu A., Ruda M., and Mlczoch J. (1998). Randomized, double-blind study comparing saruplase with streptokinase therapy in acute myocardial infarction: The compass equivalence trial. *Journal of the American College of Cardiology*, 31:487–493.

Upton G. J. G. (1982). A comparison of alternative tests for the 2x2 comparative trial. *Journal of the Royal Statistical Society Series A*, 145:86–105.

Lebenslauf

Name	Guido Skipka
Geburtstag	21.12.1969
Geburtsort	Simmerath (Kreis Aachen)
1976 - 1980	Grundschule am Schwarzwasser in Ahe (Erftkreis)
1980 - 1989 Juni 1989	Erft-Gymnasium in Bergheim (Erftkreis) Abitur
1989 - 1990	Grundwehrdienst als Sanitäter in Koblenz
1990 - 1997	Studium der Statistik mit Nebenfach Medizin an der Universität Dortmund
November 1997	Diplomprüfung, Thema der Diplomarbeit: "Der Einfluß von Ausreißern auf den einseitigen nichtzentralen t-Test im Parallelgruppendesign" Betreuer: Prof. Dr. H.J. Trampisch und Prof. Dr. S. Schach
März 1994 - Dez. 1997	studentische Hilfskraft an der Abteilung für Medizinische Informatik, Biometrie und Epidemiologie, Ruhr-Universität Bochum
Jan. 1998 - Dez. 2002	wissenschaftlicher Mitarbeiter an der Abteilung für Medizinische Informatik, Biometrie und Epidemiologie, Ruhr-Universität Bochum
seit Januar 2003	wissenschaftlicher Mitarbeiter am Institut für Mathematische Stochastik der Georg-August-Universität Göttingen