Efficient experimental design for the Behrens-Fisher problem with application to bioassay

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
holger.dette@ruhr-uni-bochum.de

Timothy E. O'Brien
Loyola University Chicago
Department of Mathematics and Statistics
6525 N. Sheridan Road
Chicago, IL 60626 USA

August 4, 2003

Abstract

A common approach in the design of experiment for the problem of comparing two means from a normal distribution is to assume knowledge of the ratio of the population variances. The optimal sampling ratio is proportional to the square root of this quantity. In this paper it is demonstrated that a misspecification of this ratio can cause a substantial loss in power of the corresponding tests. As a robust alternative a maximin approach is used to construct designs, which are efficient, whenever the experimenter is able to specify a specific region for the ratio of the population variances. The advantages of the robust designs for inference in the Behrens-Fisher problem are illustrated by means of a simulation study and an application to the design of experiment for bioassay is presented.

Keywords and Phrases: Behrens-Fisher problem, bioassay, design of experiment, local optimal design, robust designs.

1 Introduction

The problem of comparing the means of two populations is of fundamental importance in applied statistics. Let μ_i, σ_i^2 denote the population mean and variance of the ith population (for $i = 1, 2$) then the parameter of interest is typically the difference of the means $\mu = \mu_1 - \mu_2$ or the ratio $\rho = \mu_2/\mu_1$. If the ratio $\kappa = \frac{\sigma_2^2}{\sigma_1^2}$ of the population variances is unknown and the assumption of a normal distribution is made, the scenario is called Behrens-Fisher problem [see Scheffé (1970)]. There is a large number of papers in which various tests are suggested concerning hypothesis regarding the difference of the means μ. In the case of testing simple hypotheses, Welch's approximate t-solution [see Welch (1936, 1938)] appears to be a good compromise between a test which is unbiased on one hand and which is practical on the other; see for example Wang (1971) and Best and Rayner (1987). This approach was further extended in Dannenberg, Dette and Munk (1994) for testing interval hypotheses.

In contrast to the goal of constructing useful tests for the Behrens-Fisher problem, the problem of allocating observations to both populations if the total sample size has been fixed has not found much attention in the literature. It is well known [see e.g. Staudte and Sheater (1990)] that if n_1 and n_2 denote the sample sizes of both populations, the power of Welch's test is maximized if $n_1/n_2 \approx \kappa^{-1/2} = \sigma_1/\sigma_2$. A similar observation was made by Dannenberg, Dette and Munk (1994) in the context of testing interval hypotheses of the form $H_0 : \mu \notin [-\Delta, \Delta], H_1 : \mu \in [-\Delta, \Delta]$. However, these results are "local" in the sense of Chernoff (1953) as they require knowledge of the population variances in order to determine n_1 and n_2. Consequently, a misspecification of κ can yield a substantial loss in power if the sample sizes are chosen according to the rule $n_1/n_2 \approx \kappa^{-1/2}$.

We demonstrate in Section 2 by means of a simulation study that the loss of power caused by such a misspecification can be substantial. As an alternative, we propose the maximization of the minimum of an appropriately standardized power function (taken over a certain range for the parameter κ) with respect to the proportion of total observations in the first sample. We also give an explicit formula for the relative proportions for both samples with respect to the new criterion, and we demonstrate the ease with which this technique can be applied in practical settings. It is demonstrated by means of a simulation study that the new designs are robust and efficient whenever a range for the unknown ratio of the population variances can be specified.

Our new methodology is applied to the classical problem of testing the difference of two normal means and to the important problem of inference about the ratio of these means useful in direct bioassays.

2 Local optimal allocation of sample sizes

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ denote two independent samples of i.i.d. observations such that $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ $(i = 1, \ldots, n_1)$; $Y_j \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $(j = 1, \ldots, n_2)$ and consider the one-sided problem of testing the hypotheses

$$H_0 : \mu := \mu_1 - \mu_2 \leq 0 \quad \text{versus} \quad H_1 : \mu > 0. \tag{2.1}$$

In a famous paper, Welch (1938) suggested the rejection of the null hypothesis if

$$\frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{1}{n_1}\widehat{S}_1^2 + \frac{1}{n_2}\widehat{S}_2^2}} > t_{1-\alpha,\widehat{f}}, \tag{2.2}$$

where $\bar{X}_{n_1}, \bar{Y}_{n_2}$ denote sample means, $\widehat{S}_1^2, \widehat{S}_2^2$ are the common estimators of the variance and $t_{1-\alpha,\widehat{f}}$ is the quantile of the t-distribution with

$$\widehat{f} = \frac{\left(\frac{\widehat{S}_1^2}{n_1} + \frac{\widehat{S}_2^2}{n_2}\right)^2}{\left(\frac{\widehat{S}_1^2}{n_2}\right)^2/(n_1-1) + \left(\frac{\widehat{S}_2^2}{n_2}\right)^2/(n_2-1)} \tag{2.3}$$

(estimated) degrees of freedom. It was pointed out by Scheffé (1970) and Wang (1971) that this test provides a good compromise between tests which should on the one hand be unbiased and on the other hand be easily implemented. Further, it is well known [see Staudte and Sheater (1990), p. 180] that for local alternatives of the form

$$\mu = \frac{\sigma_1}{\sqrt{n_1+n_2}} \tag{2.4}$$

the asymptotic power function of this test is given by

$$\pi(\kappa) = \Phi\Big(\Big\{\frac{1}{w} + \frac{\kappa}{1-\omega}\Big\}^{-1/2} - u_{1-\alpha}\Big) \tag{2.5}$$

where $\kappa = \sigma_2^2/\sigma_1^2$ is the ratio of the population variances, $u_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the quantile of the standard normal distribution and

$$w = \lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \frac{n_1}{n_1+n_2} \in (0,1) \tag{2.6}$$

is the relative proportion of total observations in the first sample. It was pointed out by Dette and Munk (1997) that $\pi(\kappa)$ also coincides with the asymptotic power function of the extension of Welch's test to the problem of testing the equivalence hypotheses

$$H_0 : \mu \notin [-\Delta, \Delta]; H_1 : \mu \in [-\Delta, \Delta] \tag{2.7}$$

under contiguous alternatives $\mu = \Delta + \sigma_1(n_1+n_2)^{-1/2}$. A simple calculation shows that the power $\pi(\kappa)$ is maximal if

$$\frac{n_1}{n_1+n_2} \approx w_\kappa^* := \frac{1}{1+\sqrt{\kappa}} = \frac{1}{1+\sigma_2/\sigma_1}, \tag{2.8}$$

and we will call ω_κ^* the local optimal design [for testing the hypotheses (2.1) or (2.7)]. The phrase "local" is due to Chernoff (1953) and used because the optimal allocation to both samples depends on the unknown parameter $\kappa = \sigma_2^2/\sigma_1^2$. If some information regarding the ratio of population variances is available, the power of Welch's test can be increased substantially by using the rule (2.8). However, the following example shows that in general the local optimal design is indeed sensitive with respect to misspecification of the parameter κ.

Example 2.1. We have conducted a small simulation study, where $\mu = 1, \sigma_1^2 + \sigma_2^2 = 5$ and the "true" ratio $\kappa_t^{1/2} = \sigma_2/\sigma_1$ varies between 1 and 1/5. We have calculated the rejection probabilities of Welch's test (2.2) with nominal level 5% for the hypotheses (2.1) for various designs, which are calculated under the respective assumptions that the ratio is given by $\kappa_a^{1/2} = 1, 1/3, 1/5$. In other words, if $\kappa_t \neq \kappa_a$ the design was calculated under a misspecification for the ratio of the population variances. The local optimal designs are obtained by a simple rounding procedure from the values $w_\kappa^*(n_1+n_2) = (n_1+n_2)(1+\sqrt{\kappa})^{-1/2}$, which gives the sample size for the first sample. The rejection probabilities of the test (2.2) are calculated by 10,000 simulation runs, while the total sample sizes is $n_1 + n_2 = 25$ or $n_1 + n_2 = 50$.

	$n_1+n_2=25$				$n_1+n_2=50$			
$\kappa_a^{1/2}$	1	1/3	1/5	robust	1	1/3	1/5	robust
$\kappa_t^{1/2}$	$n_1=13$ $n_2=12$	$n_1=19$ $n_2=6$	$n_1=21$ $n_2=4$	$n_1=17$ $n_2=8$	$n_1=25$ $n_2=25$	$n_1=37$ $n_2=13$	$n_1=41$ $n_2=9$	$n_1=33$ $n_2=17$
1	0.444	0.350	0.269	0.406	0.715	0.581	0.448	0.669
1/3	0.448	0.519	0.483	0.521	0.694	0.785	0.770	0.784
1/5	0.451	0.573	0.577	0.537	0.697	0.830	0.838	0.798

Table 2.1: *Rejection probabilities of Welch's test (2.2) for the hypotheses (2.1) for various designs and ratios $\kappa_t = \sigma_2^2/\sigma_1^2$ of population variances.*

The differences between the rejection probabilities are remarkable. For example, if the "true" ratio of the population variances is given by $\kappa_t^{1/2} = 1$, but the local optimal design is found under the assumption that $\kappa_a^{1/2} = 1/3$. then we observe for the sample size $n_1+n_2 = 50$ a loss of power of approximately 19%. The results indicate that the optimal allocation rule (2.8) is rather sensitive with respect to a misspecification of the unknown ratio of the population variances. In the fourth columns (labeled "robust"), the table also contains a design which is both quite robust and efficient for all situations under consideration. For example, if $n_1 + n_2 = 25$ the loss of efficiency of the allocation rule $n_1 = 17, n_2 = 8$ compared to the best design is only approximately 9% ($\kappa_t^{1/2} = 1$), 0.4% ($\kappa_t^{1/2} = 1/3$) and 7% ($\kappa_r^{1/2} = 1/5$). By comparison, the allocation rule $n_1 = 19, n_2 = 6$ (corresponding to the assumption $\kappa_a^{1/2} = 1/3$) yields a loss of efficiency of 21% ($\kappa_t^{1/2} = 1$) and 1% ($\kappa_t^{1/2} = 1/5$) while it is the best for $\kappa_t = 1/3$. Similarly, the loss of efficiency of the allocation rule $n_1 = 21$ $n_2 = 4$ (corresponding to the assumption $\kappa_a^{1/2} = 1/5$) is approximately 39% ($\kappa_t^{1/2} = 1$) and 7% ($\kappa_t^{1/2} = 1/3$).

The robust designs were calculated by a maximin approach which will be developed in the following section, and which uses only the information that the ratio of the population standard deviations lies in the interval [1/5, 1]. We feel this is the more realistic setting since practitioners will rarely be able to give an accurate point estimate for the ratio of the variances, whereas an accurate interval estimate can usually be given. The results of Table 2.1 along with additional simulations (not shown for the sake of brevity) indicate that robust and efficient designs are available if an interval for the unknown ratio of the population variances can be specified by the experimenter.

3 Robust designs for the Behrens-Fisher problem

Note that the power function of the test (2.2) increases with the expression

$$f(w,\kappa) = \Big\{\frac{1}{w} + \frac{\kappa}{1-w}\Big\}^{-1} \tag{3.1}$$

and that thus the locally optimal design $w_\kappa^* = 1/(1+\kappa^{1/2})$ is found by maximizing $f(w,\kappa)$ with respect to w for given κ. In Example 2.1, we indicated that these designs are not necessarily robust with respect to a misspecification of the unknown ratio of population variances. For the construction of a more robust design, we assume that an interval, say $[\kappa_L, \kappa_U]$, for the unknown population variance can be specified by the experimenter, and consider the efficiency

$$\text{eff}(w,\kappa) = \frac{f(w,\kappa)}{\max_v f(v,\kappa)} = \frac{(1+\sqrt{\kappa})^2}{\frac{1}{w} + \frac{\kappa}{1-w}}. \tag{3.2}$$

Note that the efficiency, which varies between 0 and 1, measures the performance of the design w with respect to the best design provided κ is the "true" ratio of the population variances. A design w^* is called standardized maximin optimal if it maximizes the minimum efficiency

$$g(w) = \min_{\kappa\in[\kappa_L,\kappa_U]} \text{eff}(w,\kappa) \tag{3.3}$$

over the interval $[\kappa_L, \kappa_U]$. This design criterion is similar to the standardized optimality criteria used in Dette (1997) and Imhof (2001). Further, it is established in the Appendix that for fixed w the function $\kappa \to \text{eff}(w,\kappa)$ is unimodal with at most one maximum in the interval $[\kappa_L, \kappa_U]$ (see Lemma A.1). It therefore follows that

$$g(w) = \min\{\text{eff}(w,\kappa_L), \text{eff}(w,\kappa_U)\}. \tag{3.4}$$

Moreover, in Lemma A.2 in the Appendix, we show that for the standardized maximin optimal design

$$w^* = \arg\max_{w\in[0,1]} g(w)$$

it follows that $\text{eff}(w^*,\kappa_L) = \text{eff}(w^*,\kappa_U)$. This equality determines the optimal design as

$$w^* = \frac{2+{\kappa_L}^{1/2}+{\kappa_U}^{1/2}}{2(1+{\kappa_L}^{1/2})(1+{\kappa_U}^{1/2})} \tag{3.5}$$

for which the minimal efficiency is

$$g(w^*) = \frac{(2+{\kappa_L}^{1/2}+{\kappa_U}^{1/2})\{{\kappa_L}^{1/2}(1+{\kappa_U}^{1/2}) + {\kappa_U}^{1/2}(1+{\kappa_L}^{1/2})\}}{2(1+{\kappa_L}^{1/2})(1+{\kappa_U}^{1/2})({\kappa_L}^{1/2}+{\kappa_U}^{1/2})}. \tag{3.6}$$

Example 3.1. For the situation considered in Example 2.1, we have ${\kappa_L}^{1/2} = 1/5$ and ${\kappa_U}^{1/2} = 1$, which yields the standardized maximin optimal design weight $w^* = 2/3$ and for which the minimal efficiency is $g(w^*) = 8/9$. This high value of the minimal value of the design efficiency underscores the remarkable robustness of our robust design. Incidentally, the corresponding (potentially non-rational) weight is translated into a practical design allocation for the first sample by rounding $(n_1 + n_2) \cdot w^* = (n_1 + n_2)2/3$ to the nearest integer (as in Table 2.1).

Remark 3.2. We also note that the design problem is symmetric in the following sense. If $w^*_{\kappa_L,\kappa_U}$ denotes the (standardized maximin) optimal proportion for the first sample if the parameter κ is assumed to be in the interval $[\kappa_L, \kappa_U]$, then the corresponding quantity for the interval $[1/\kappa_U, 1/\kappa_L]$ satisfies

$$w^*_{1/\kappa_U,1/\kappa_L} = 1 - w^*_{\kappa_L,\kappa_U}.$$

In other words, the standardized maximin optimal design for the interval $[1/\kappa_U, 1/\kappa_L]$ can be obtained from the corresponding design for the interval $[\kappa_L, \kappa_U]$ by interchanging the role of the sample sizes n_1 and n_2. For this reason the robust designs can easily be tabulated. Some designs for selected values of κ_L and κ_U are presented in Table 3.1 for the sake of completeness. Finally, we note that this symmetry implies that the equal allocation rule $w^* = 1/2$ is (standardized maximin) optimal for any interval of the form $[1/\kappa_0, \kappa_0]$ where $\kappa_0 > 1$.

Example 3.3. The results derived so far have been derived under the assumption that one-sided hypotheses are tested with Welch's approximate t-solution. It follows from Dette and Munk (1997) that these results are directly applicable to the problem of testing the equivalence hypotheses $H_0 : \mu \notin [-\Delta, \Delta]$; $H_1 : \mu \in [-\Delta, \Delta]$, because the asymptotic power function coincides with that of the one-sided problem.
In principle, a similar analysis could be performed for cases where simple hypotheses $H_0 : \mu = 0; H_1 : \mu \neq 0$ or interval hypotheses $H_0 : \mu \in [-\Delta, \Delta]; H_1 : \mu \notin [-\Delta, \Delta]$ are of interest. However, our numerical results show that the designs derived for the one-sided problem are also very efficient for testing other hypotheses. By way of illustration, consider the situation of Example 2.1 where $\sigma_1^2+\sigma_2^2 = 5$ and a test with level 5% for the hypotheses $H_0 : \mu = 0; H_1 : \mu \neq 0$ has to be performed. In order to demonstrate the application of Remark 3.2, we consider the cases where $\kappa_t^{1/2} = 1, 3, 5$ for the true value of the ratio of the variances, while we assumed ${\kappa_L}^{1/2} = 1$ and ${\kappa_U}^{1/2} = 5$ for the construction of the robust design. The optimal proportion for the first sample is now given by $w^* = 1/3$ and the simulated rejection probabilities are given in Table 3.2 for sample sizes $n_1 + n_2 = 25$ or 50. We observe a similar picture as for the one-sided case. The local optimal designs are sensitive with respect to misspecification of the unknown ratio of population variances, while the standard maximin optimal designs yield a reasonable power in all cases under consideration.

$\kappa_L \backslash \kappa_U$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	0.760	0.725	0.703	0.686	0.673	0.662	0.652	0.644	0.636	0.630
0.2		0.691	0.668	0.652	0.638	0.627	0.618	0.609	0.602	0.595
0.3			0.646	0.629	0.616	0.605	0.595	0.587	0.580	0.573
0.4				0.613	0.599	0.588	0.579	0.570	0.563	0.556
0.5					0.586	0.575	0.565	0.557	0.549	0.543
0.6						0.563	0.554	0.546	0.538	0.532
0.7							0.544	0.536	0.529	0.522
0.8								0.528	0.521	0.514
0.9									0.513	0.507
1.0										0.5

Table 3.1. *Standardized maximin optimal designs for various intervals $[\kappa_L, \kappa_U]$ for the unknown ratio $\kappa = \sigma_2^2/\sigma_1^2$ of the population variances. The value w^* in the table gives the relative proportion of total observations in the first sample.*

	$n_1 + n_2 = 25$				$n_1 + n_2 = 50$			
$\kappa_a^{1/2}$	1	3	5	robust	1	3	5	robust
$\kappa_t^{1/2}$	$n_1 = 12$ $n_2 = 13$	$n_1 = 6$ $n_2 = 19$	$n_1 = 4$ $n_2 = 21$	$n_1 = 8$ $n_2 = 17$	$n_1 = 25$ $n_2 = 25$	$n_1 = 12$ $n_2 = 38$	$n_1 = 8$ $n_2 = 42$	$n_1 = 17$ $n_2 = 33$
1	0.322	0.256	0.206	0.283	0.593	0.460	0.355	0.538
3	0.301	0.398	0.385	0.377	0.587	0.689	0.665	0.674
5	0.299	0.418	0.433	0.399	0.575	0.730	0.753	0.695

Table 3.2. *Rejection probabilities of Welch's test of a simple hypothesis for various designs and ratios $\kappa_t = \sigma_2^2/\sigma_1^2$ of population variances.*

4 Application to Bioassay

One concern of bioassay, or biological assays, is the estimation of the potency of one drug (B) relative to another (A), typically involving comparing a new drug with a standard. Further, in contrast with indirect assays, direct assays hold that the necessary concentrations that produce the same therapeutic effect can be directly measured. In this setting, the relative potency (ρ) of drug B to A is the ratio of the respective means, where the underlying respective distributions are assumed to be Gaussian $A \sim \mathcal{N}(\mu_1, \sigma_1^2), B \sim \mathcal{N}(\mu_2, \sigma_2^2)$; thus, $\rho = \mu_2/\mu_1$. Further background of direct assays is given in Finney (1978, Ch. 2) and Govindarajulu (2000, Ch. 2).

Often practitioners are interested in a confidence interval for the relative potency, and experimental designs which produce shorter confidence intervals are therefore desired. In the case of independent populations, a standard calculation shows that the first order approximation for the length of any reasonable confidence interval is proportional to the root of the function

$$g(w,\kappa,\rho) = \frac{1}{w} + \frac{\kappa/\rho^2}{1-w},$$

and all results of the previous sections are therefore applicable to this case but with κ replaced by κ/ρ^2. For example, the local optimal design uses

$$w^*_{\kappa/\rho} = \frac{1}{1+\sqrt{\kappa}/\rho} \tag{4.1}$$

as the weight for the first sample. Similarly, if the experimenter is able to specify a region, say $[\kappa_L, \kappa_U]$ for the quantity κ/ρ^2 the optimal design is given by (3.5).

Consider for example the situation where the population variances are the same, i.e. $\kappa = 1$, and a confidence interval is constructed using Fieller's theorem [Finney (1978)]. This interval is of the form

$$\rho_{L,U} = \Big[\widehat{\rho} \pm \frac{t\widehat{s}}{\bar{X}_{n_1}}\Big\{\frac{1}{n_2} + \widehat{\rho}^2\frac{1}{n_1} - \frac{g}{n_2}\Big\}^{1/2}\Big]/(1-g)$$

where $g = t^2s^2/(n_1\bar{X}^2_{n_1})$, t is the $(1-\alpha)$ quantile of the t-distribution with n_1+n_2-2 degrees of freedom, $\widehat{\rho} = \bar{Y}_{n_2}/\bar{X}_{n_1}$ and $\bar{S}^2$ is the pooled variance estimate. To highlight the benefits of our robust design strategy, we have performed a small simulation study to calculate the average length

$$\widehat{L} = \rho_U - \rho_L$$

of this interval for different designs. For this simulation, the true relative potency ρ_t varies between 1, 2.25, 4 and 6.25, and for the construction of the locally optimal designs by formula (4.1) we again assume $\sigma_1^2 = \sigma_2^2 = 0.25$, (whence $\kappa = 1$). The results are given in Table 4.1 and show that the length depends strongly on the specification of the relative potency. Thus, a misspecification of this quantity can produce a substantially larger confidence interval. For example, if the true relative potency is $\rho_t = 1$ but we use a design based on the assumption $\rho_a = 4$, the length of the resulting confidence interval is increased by 23 %. On the other hand, the robust design given in the table is constructed under the assumption that the true ρ_t lies in the interval $[1, 6.25]$, and yields the optimal weight $w^* = 0.607$ (using formula (3.5)). For the total sample size $n_1 + n_2 = 50$, this weight translates into the allocation $n_1 = 30$ and $n_2 = 20$, for a total sample size of $n_1+n_2 = 50$. From equation (3.6), this robust design has an efficiency of at least 95.41% . This fundamental result is illustrated in our simulation study, which shows that the robust design is indeed both robust to the choice of ρ and very efficient with a loss of efficiency of at most 5% (see Table 4.1).

	$n_1 + n_2 = 50$				
ρ_a	1.0	2.25	4.0	6.25	robust
	$n_1 = 25$	$n_1 = 35$	$n_1 = 40$	$n_1 = 43$	$n_1 = 30$
ρ_t	$n_2 = 25$	$n_2 = 15$	$n_2 = 10$	$n_2 = 7$	$n_2 = 20$
1.0	0.493	0.529	0.605	0.694	0.498
2.25	0.632	0.624	0.676	0.750	0.614
4.0	0.791	0.736	0.770	0.829	0.748
6.25	0.954	0.856	0.869	0.921	0.887

Table 4.1. *Simulated length of the confidence interval for the relative potency based on Fieller's theorem for various designs and different values of* $\rho_t = \mu_2/\mu_1$.

5 Concluding remarks

In this paper we have determined efficient and robust designs for Welch's approximate t-test for testing one-sided hypotheses. Our method is based on a maximin approach and we have shown their usefulness and superiority in the setting of both the classical difference of two means and for the relative potency of similar compounds. An explicit formula for the proportions of total observations for both samples is given and the designs can easily be implemented if the experimenter is able to specify a region $[\kappa_L, \kappa_U]$ for the unknown ratio $\kappa = \sigma_2^2/\sigma_1^2$ of the population variances. It is demonstrated by means of a simulation study that the derived designs yield to an efficient inference for all $\kappa \in [\kappa_L, \kappa_U]$, whenever $0.2 \leq {\kappa_L}^{1/2} \leq {\kappa_U}^{1/2} \leq 1$ (equivalently $1 \leq {\kappa_L}^{1/2} \leq {\kappa_U}^{1/2} \leq 5$). This should encompass most cases of practical interest. An experiment with a larger (smaller) ratio of standard deviations should never be performed because the power of the Welch test becomes very small.

We have concentrated on one-sided hypotheses of the form (2.1) for the sake of brevity. However, for the problem of testing the equivalence hypotheses $H_0 : \mu \notin [-\Delta, \Delta]$; $H_1 : \mu \in [-\Delta, \Delta]$ it is shown in Dette and Munk (1997) that the asymptotic power function of an extension of Welch's test coincides with the power function of the test for one-sided hypotheses. As a consequence the results obtained in this paper are applicable for testing interval hypotheses by Welch's approximate t-solution introduced by Dannenberg, Dette and Munk (1994). Moreover, it is demonstrated that the designs derived in Section 3 also provide a robust and efficient allocation for the problem of testing simple hypotheses. For these reasons we recommend to use these designs for the Behrens-Fisher problem of testing the difference of two means whenever an interval for the ratio of the population variances can be specified.

The results are also applicable for the classical problem of bioassay where the goal of the experiment is the estimation of the potency of one drug relative to another. For this problem, robust and efficient designs can be obtained from the results of this paper whenever the experimenter is able to specify an interval for the ratio κ/ρ^2 where ρ is the unknown relative potency and κ the ratio of the population variances.

6 Appendix

Lemma A.1. *For fixed w the function $\kappa \to$ eff(w, κ) defined in (3.2) is unimodal with at most one maximum in the interval $[\kappa_L, \kappa_U]$.*

Proof. Recall the definition of the efficiency in (3.2). A straightforward calculation shows that

$$\frac{\partial}{\partial\widetilde{\kappa}}\Big(\log(\text{eff}(w, \widetilde{\kappa}^2)\Big) = 2\frac{(\widetilde{\kappa}+1)w-1}{(1+\widetilde{\kappa})(w-1-w\widetilde{\kappa})},$$

which vanishes only at the point $\widetilde{\kappa} = (1-w)/w$. A similar calculation of the second derivative yields

$$\frac{\partial^2}{\partial^2\widetilde{\kappa}}\log(\text{eff}(w, \widetilde{\kappa}^2)\Big|_{\widetilde{\kappa}=\frac{(1-w)}{w}} = \frac{2w^3}{(w-1)(w+(1-w))^2} < 0.$$

Consequently it follows that the function eff$(w, \widetilde{\kappa})$ has at most one extremum in the interval $[\kappa_L, \kappa_U]$, which is a maximum. $\Box$

Lemma A.2. *If $w^*_{\kappa_L,\kappa_U}$ denotes the standardized maximin optimal design, then*

$$\text{eff}(w^*_{\kappa_L,\kappa_U}, \kappa_L) = \text{eff}(w^*_{\kappa_l,\kappa_U}\kappa_U).$$

Proof. We can split the maximization of the right hand side of (3.4) in the maximization over the sets

$$\begin{aligned}
\mathcal{M}_< &= \Big\{w \in [0,1] \mid \text{eff}(w, \kappa_L) < \text{eff}(w, \kappa_U)\Big\},\\
\mathcal{M}_> &= \Big\{w \in [0,1] \mid \text{eff}(w, \kappa_L) > \text{eff}(w, \kappa_U)\Big\},\\
\mathcal{M}_= &= \Big\{w \in [0,1] \mid \text{eff}(w, \kappa_L) = \text{eff}(w, \kappa_U)\Big\}.
\end{aligned}$$

Now assume that $w^*_{\kappa_L,\kappa_U} \in \mathcal{M}_<$. In this case we obtain $w^*_{\kappa_L,\kappa_U} = 1/(1+\sqrt{\kappa_L})$ and by the definition of $\mathcal{M}_<$ the inequality

$$\text{eff}(\frac{1}{1+\sqrt{\kappa_L}}, \kappa_L,) < \text{eff}(\frac{1}{1+\sqrt{\kappa_L}}, \kappa_U).$$

But this inequality is equivalent to

$$(\sqrt{\kappa_L} - \sqrt{\kappa_U})^2 < 0,$$

which yields a contradiction. A similar argument for the set $\mathcal{M}_>$ shows that the maximum is attained in $\mathcal{M}_=$, which completes the proof. $\Box$

Acknowledgements. The authors gratefully acknowledge Isolde Gottschlich, who typed this paper with considerable technical expertise, and the Deutsche Forschungsgemeinschaft (SFB 475: Komplexitätsreduktion in multivariaten Datenstrukturen) for financial support.

References

Best, D.J. & Rayner, J.C.W. (1987). Welch's approximate solution for the Behrens-Fisher problem. Technometrics **29**, 205-10.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. Ann. Math. Statist. **24**, 586-602.

Dannenberg, O., Dette, H. & Munk, A. (1994). An extension of Welch's approximate t-solution to comparative bioequivalence trials.
Biometrika **81**, 91-101.

Dette, H. (1997). Designing experiments with respect to 'standardized' optimality criteria. J. Roy. Statist. Soc. Ser. B **59**, 97-110.

Dette, H. & Munk, A. (1997). Optimum allocation of treatments for Welch's test in equivalence assesment. Biometrics **53**, 1143-1150.

Finney, D. J. (1978). Statistical Method in Biological Assay, 3rd edition, London: Charles Griffin & Co.

Govindarajulu, Z. (2000). Statistical Techniques in Bioassay, 2nd edition, Basel: Karger.

Imhof, L.A. (2001). Maximin designs for exponential growth models and heteroscedastic polynomial models. Ann. Statist. **29**, 561-576.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. J. Am. Statist. Assoc. **332**, 1501-8.

Staudte, R.G. & Sheater, S.J. (1990). Robust Estimation and Testing. New York: John Wiley.

Wang, Y.Y. (1971). Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem. J. Am. Statist. Assoc. **66** (33), 605-8.

Welch, B.L. (1936). Specification of rules for rejecting the variable, a product with particular reference to an electric lamp problem. J.R. Statist. Soc. Suppl. **3**, 29-48.

Welch, B.L. (1938). The significance of the difference between means when the population variances are unequal. Biometrika **29**, 350-62.