

Interim Analysis and Sample Size Reassessment

Martin Posch

*Peter Bauer**

Keywords: Sample size reassessment, Adaptive two stage designs, Mid-trial modification, Combination Tests, Conditional Power

2000

*Both: Department of Medical Statistics, University of Vienna, Schwarzschanerstr. 17, A-1090 Vienna, Austria, email: Martin.Posch@univie.ac.at, Peter.Bauer@univie.ac.at

Abstract

This paper deals with the reassessment of the sample size for adaptive two stage designs based on conditional power arguments utilizing the variability observed at the first stage. Fisher's product test for the p -values from the disjoint samples at the two stages is considered in detail for the comparison of the means of two normal populations. We show that stopping rules allowing for the early acceptance of the null hypothesis which are optimal with respect to the average sample size may lead to a severe decrease of the overall power if the sample size is a priori underestimated. This problem can be overcome by choosing designs with low probabilities of early acceptance or by mid-trial adaptations of the early acceptance boundary using the variability observed in the first stage. This modified procedure is negligibly anti-conservative and preserves the power.

1 Introduction

The design of experiments is generally based on more or less restrictive assumptions in the planning phase. For example designs that are optimal with respect to sample size under specific assumptions may be poor if the assumptions do not apply as in the case where the variability is misjudged a priori. A flexible tool to address this issue are adaptive designs. Adaptive two stage designs expand the notion of an internal pilot study such that the design of the remaining study part can be modified without compromising the overall significance level or the overall power of the trial (Bauer and Köhne, 1994; Bauer and Röhmel 1995).

In this paper we investigate the situation where in the planning phase of an experiment the power is fixed at a minimum relevant effect size for a given type I error probability. This effect size worth to be detected in the trial is assumed to be

given in absolute differences of population means which is commonly the case. To calculate the sample size usually a variance estimate is taken from previous data. However, there may be considerable differences in the environment (population, measurement procedures, treatment modalities, etc) which may influence the variability in the forthcoming trial. This – in case of an underestimation of the variability in the planning phase – can lead to a considerable loss of power. There seems to be a tendency in medical experiments to use optimistically low variance estimates to keep the number of patients low. However, Kieser and Friede (1999) also give an example for an overestimation in the planning phase which leads to an unnecessarily high sample size.

A way out is to use an internal estimate of the variability for the reassessment of sample size in the course of the study. Several proposals for sample size reassessment after an internal pilot study have been made (Stein, 1945; Coffey and Muller, 1999; Wittes et al., 1999; Zucker et al., 1999; Kieser and Friede, 1999). If in the final analysis the common test statistics is applied without accounting for the reassessment procedure (Wittes and Brittain, 1990; Birkett and Day, 1994) type I error probability inflations have been shown. These can be of relevant magnitude in particular if the preplanned sample size may be reduced in the reassessment procedure.

This paper deals with procedures where decision rules for early rejection and acceptance of the null hypothesis are applied after the first stage (internal pilot study) together with systematic sample size reassessment based on conditional power arguments. Our tools are adaptive combination tests which can handle both issues simultaneously without inflating the type I error probability.

A well studied combination test is based on Fisher's product criterion for the

one sided p -values p_1 and p_2 calculated for the tests in the disjoint samples before and after the adaptive interim analysis. The level α combination test rejects if

$$p_1 p_2 < c_\alpha = e^{-0.5 \chi_{4,1-\alpha}^2} \quad (1)$$

where $\chi_{n,1-\alpha}^2$ denotes the $1-\alpha$ quantile of the central χ^2 -distribution with n degrees of freedom (we assume that under the null hypothesis p_1 and p_2 are independently and uniformly distributed on $[0, 1]$). There is an appealing feature of this method: All information from the first stage (e.g. the observed variability and effect size) can be used to plan the second stage. Sequential procedures for early stopping with acceptance and rejection of the null hypothesis can be easily defined (see Section 2.2.1 below).

Recently several other combination tests have been suggested for the adaptive scenario. For normally distributed test statistics with known variance Lehman and Wassmer (1999) use the sum of the standardized test statistics (see also Bauer and Köhne, 1994). This is an application of Lipták's (1958) combination test statistics $\Phi^{-1}(1 - p_1) + \Phi^{-1}(1 - p_2)$. Shen and Fisher's (1999) approach restricted to two stages corresponds to a weighted sum of the two inverse normal terms (Mosteller and Bush, 1954). The method of Cui, Hung and Wang (1999) also applies a combination principle with prefixed weights. The setup of Proschan and Hunsberger (1995) is based on the concept of the conditional error function. Combination tests can be looked at in terms of the error function and vice versa (Posch and Bauer, 1999; Wassmer, 1999). Conditional power arguments for sample size reassessment can be applied to all the two stage decision procedures. However, due to its generality, simplicity and high flexibility we only consider the product test further on. Its power appears to be very close to the power of e.g. the method of Proschan and Hunsberger (1995) for normally distributed

statistics (Wassmer, 1998).

The paper is structured in four steps. First we describe the sample size re-assessment procedure for adaptive two stage designs based on conditional power. Then we derive some asymptotic results on optimal decision boundaries in terms of the average sample size when systematic sample size reassessment is applied. Further we investigate the statistical properties of these designs for the finite case when the choice of optimal designs has been based on a correct or incorrect a-priori specification of the variance. Finally, to overcome a possible loss of power in case of an a priori underestimation of the variability we propose a mid-trial redesigns of the decision boundaries.

2 Reassessment of sample size

2.1 The test problem

We consider a two stage test of the one sided hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ for the difference of means from two independent normal populations and assume common unknown variance σ^2 in both groups. Let n_1, n_2 denote the sample sizes in each treatment group for the first resp. second stage (for simplicity we assume that the sample sizes are balanced over treatments).

2.2 A general reassessment procedure based on conditional power including early stopping

2.2.1 Stopping rules

General stopping rules can be defined in a straightforward way. If $p_1 < \alpha_1$ we stop early with a rejection of the null hypothesis tested at the first stage; if $p_1 \geq \alpha_0$ we stop early with the acceptance of the null hypothesis after the first stage. Only if $\alpha_1 \leq p_1 < \alpha_0$ we do proceed to the second stage. Here we reject if $p_1 \cdot p_2 < c_{\alpha_2}$, otherwise we accept. To get a level α test from two independent uniformly distributed random variables the quantities α_0 , α_1 and α_2 have to satisfy (Bauer, Bauer and Budde, 1998) :

$$\alpha_1 + c_{\alpha_2} \{\log(\alpha_0) - \log(\alpha_1)\} = \alpha, \quad (2)$$

where c_{α_2} is defined in (1) replacing α by α_2 .

Feasible choices of α_0 and α_1 Since $p_1 p_2 \leq p_1$ it only makes sense to consider cases where $\alpha_1 \geq c_{\alpha_2}$. By (2) this is the case if $\alpha_0 \geq \alpha_1 e^{\frac{\alpha - \alpha_1}{c_{\alpha_2}}}$. For fixed α_2 we can write α_0 as function of α_1

$$\alpha_0(\alpha_1) = \alpha_1 e^{\frac{\alpha - \alpha_1}{c_{\alpha_2}}}. \quad (3)$$

The function $\alpha_0(\alpha_1)$ is monotonically decreasing in α_1 with $\alpha_0(c_\alpha) = 1$ and $\alpha_0(\alpha) = \alpha$. Some special choices have been discussed: $\alpha_0 < 1, \alpha_2 = \alpha$ (Bauer and Köhne, 1994); $\alpha_0 = 1$, which means no early acceptance (Bauer and Röhmel, 1995). In the following we choose $\alpha_2 \leq \alpha$ such that the final product test is always performed at a level not exceeding α . Nevertheless, the product test may reject the null hypothesis after the second stage although the fixed sample size test for the pooled samples and the same α -level does not reject. This, however, is a general feature common to adaptive designs (Posch and Bauer, 1999). Clearly

for $\alpha_0 < 1$ adherence to the acceptance boundaries is required to control the level α .

2.2.2 Sample size reassessment

The two stage procedure based on the product test provides a strikingly simple way for sample size reassessment. Let p_1 be the p -value calculated from the first stage. Then we know that the product test in the end can only reject if $p_2 < c_{\alpha_2}/p_1$. Hence, we can aim at a particular conditional power $1 - \beta$ at the end by choosing an appropriate sample size for the second stage at the modified significance level c_{α_2}/p_1 . Thus, we can use all classical sample size calculations for fixed sample size tests (Bauer and Köhne, 1994). If $\alpha_1 < p_1 < \alpha_0$ (in designs that allow early acceptance) or if $\alpha_1 < p_1$ (in designs that do not allow early acceptance), i.e. if there is a second stage, we choose n_2 to satisfy

$$n_2(p_1, \hat{\sigma}_1, \alpha_2) = \frac{2 \hat{\sigma}_1^2 \left([z_\beta + z_{\alpha^*(p_1)}]^- \right)^2}{\Delta^2}, \quad (4)$$

where $\hat{\sigma}_1^2$ is the estimated variance from the first stage, $\alpha^*(p_1) = \frac{c_{\alpha_2}}{p_1}$, z_γ denotes the γ -quantile of the standard normal distribution, and $[\cdot]^-$ denotes the negative part of the argument. Thus, we determine the sample size in the second stage based on the prefixed alternative $\Delta = \mu_1 - \mu_2 > 0$, the variance $\hat{\sigma}_1^2$ observed in the first stage, and the p -value from the first stage such that the power conditional on the event that *there is a second stage* is close to $1 - \beta$ (as long as the accuracy of the variance estimate is good). Thus, n_2 depends on p_1 (which itself depends on the estimated effect size) and on the estimated variance.

Note that asymptotically for $n_1 \rightarrow \infty$ such that $\hat{\sigma}_1 \rightarrow \sigma$ almost surely, the conditional power for the second stage is almost surely $1 - \beta$. We did not

investigate the case where upper confidence limits for the variance are used for sample size reassessment (Browne, 1995; Kieser and Wassmer, 1996) or estimates of Δ/σ (Proschan and Hunsberger, 1995). The calculations in the following can be generalized to these cases.

The normal approximation in the sample size reassessment formulas is taken for computational simplicity throughout the paper. All integrations were performed with Mathematica using a minimum of $n_2 = 3$. We approximated the t -distribution with more than 200 degrees of freedom by the normal distribution.

2.3 The procedure with early acceptance

If we allow for early acceptance we can construct a decision boundary leading to the conditional and overall power $1 - \beta$. Given a conditional power of $1 - \beta$ at the second stage, to get an overall power of $1 - \beta$, the power conditional on early stopping has also to be $1 - \beta$. Thus, under the alternative

$$\frac{P(p_1 > \alpha_0)}{P(p_1 > \alpha_0) + P(p_1 < \alpha_1)} = \beta \quad (5)$$

has to hold. From this equation we derive for given α_0 and α_1 the sample size for stage one. Let $\hat{\Delta}_1$ denote the difference between the averages of the two groups in the first sample. Then asymptotically for $\Delta \rightarrow 0$ such that $n_1 \rightarrow \infty$ the term (5) can be approximated by $\frac{P(\frac{\hat{\Delta}_1\sqrt{n_1}}{\sqrt{2}\sigma} < z_{1-\alpha_0})}{P(\frac{\hat{\Delta}_1\sqrt{n_1}}{\sqrt{2}\sigma} < z_{1-\alpha_0}) + P(\frac{\hat{\Delta}_1\sqrt{n_1}}{\sqrt{2}\sigma} > z_{1-\alpha_1})} = \beta$. Let $\xi = \frac{\Delta\sqrt{n_1}}{\sqrt{2}\sigma}$. Then ξ satisfies the equation

$$\frac{1 - \beta}{\beta} \Phi_{(\xi,1)}(z_{1-\alpha_0}) = 1 - \Phi_{(\xi,1)}(z_{1-\alpha_1}), \quad (6)$$

where $\Phi_{(\mu,\sigma)}(\cdot)$ denotes the cumulative distribution function of the normal distribution with parameters μ and σ . We denote the unique solution of (6) by a

function $\xi(\alpha_0, \alpha_1)$ which is defined for all $\alpha_1 \in (c_{\alpha_2(\alpha_0, \alpha_1)}, \alpha)$, $\alpha_0 \in (\alpha_1, 1 - \alpha_1)$, and fixed $\beta < 1/2$. Hence, if the sample size in the first stage is chosen by

$$n_1(\alpha_0, \alpha_1) = \frac{2 \hat{\sigma}_0^2 \xi(\alpha_0, \alpha_1)^2}{\Delta^2}, \quad (7)$$

where $\hat{\sigma}_0$ denotes the a priori estimated standard deviation. The global power is $1 - \beta$ if the variance estimate is correct. Note that equation (6) can be solved numerically, e.g. by Newton's method.

2.4 The procedure without early acceptance

Many sequential decision procedures do not include the possibility of a controlled early acceptance decision ($\alpha_0 = 1$). Then, since there is a positive probability to reject H_0 after the first stage the overall power is expected to be greater than $1 - \beta$ if we apply systematic sample size reassessment based on the conditional power $1 - \beta$. If an exact overall power is intended this would require that a conditional power of less than $1 - \beta$ is applied for sample size reassessment. However, if an experimenter proceeds to the second stage he or she may not be convinced to use a lower conditional power because some power has already been "spent". Therefore in this situation designs will be investigated which are expected to provide an overall power exceeding $1 - \beta$.

By the way, the design without early acceptance of the null hypothesis has drawbacks. The null hypothesis may be rejected while the overall mean (over the observations of both stages) points towards the wrong direction. An opposite trend in a large first stage may not be compensated by a clearly positive trend (leading to a very small p_2) observed in a small second stage. This may occur also for other combination tests, since the combination functions have to be laid

down a priori. When sample size reassessment occurs, the weights of the two stages in general will not reflect the actual sample sizes. In practice one could impose lower and upper limits on n_2 . This by construction of the adaptive test procedure would have no impact on the type I error probability but only increase the average power.

It may also happen that the sample size required for the second stage is larger than the sample size needed for a new trial performed at level α . However in these cases one would start the new trial with the burden of a previous negative trial.

3 Minimizing the average sample size

In this section we discuss optimal decision boundaries assuming the variance has been correctly specified a priori. We further choose the simplified setup $\Delta \rightarrow 0$ such that $n_1 \rightarrow \infty$ and $\hat{\sigma}_1 \rightarrow \sigma$ almost surely. In this situation the variability of n_2 arises from the variability of p_1 only.

3.1 The procedure with early acceptance

Consider first the case with early stopping, i.e. $\alpha_0 < 1$. For fixed α, β , and a particular set of $\alpha_1, \alpha_0, \alpha_2$ values we assume that n_1 and n_2 are determined to give a conditional power of $1 - \beta$ in case of early stopping as well as continuation based on the a priori estimate σ_0 . To get the average sample size we average over all values of p_1 . Under H_1 , if $\Delta_R = \Delta$ is the real difference, and under H_0 we can write the average sample size as $\bar{n}(\alpha_0, \alpha_1) = n_{\text{fix}} K_{H_i}(\alpha_0, \alpha_1)$, where n_{fix} denotes

the sample size of the standard test with fixed sample size and

$$K_{H_i}(\alpha_0, \alpha_1) = c^2 \left\{ \xi(\alpha_0, \alpha_1)^2 + \int_{\alpha_1}^{\alpha_0} (z_\beta + z_{\alpha^*(p)})^2 f_{H_i}(p) dp \right\},$$

where $f_{H_i}(p)$ denotes the density of p_1 under H_i and $c^2 = (z_\alpha + z_\beta)^{-2}$ (for the technical details see Appendix A). Thus, to minimize the average sample size over α_0 and α_1 under the alternative it suffices to minimize $K_{H_1}(\alpha_0, \alpha_1)$. Table 1 lists the optimal α_0 and α_1 values for different levels of α and β under the constraint $\alpha_2 = \alpha$. Minimizing under H_0 leads to similar decision boundaries. E.g. for $\alpha = 0.025$ and $\beta = 0.2$ we get the optimal $\alpha_0 = 0.203$ (in contrast to $\alpha_0 = 0.171$ if we minimize under H_1).

We investigated how much the optimal design parameters depend on the assumption of known σ and performed a numerical optimization for the actual procedure where the estimated standard deviation is used for sample size re-assessment. Here the optimal design parameters depend on the effect size. For an effect size of half of the standard deviation with $\alpha = 0.025$, $\beta = 0.2$ we get a lower optimal α_0 of 0.153. However, the average power and sample size are practically unchanged as compared to the asymptotically optimal design.

For a similar setup Case and Morgan (1987) gave optimal stopping boundaries. For their setup it makes a big difference if they optimize under H_1 or H_0 since they do not use conditional power arguments. For $\alpha = 0.01$, $\beta = 0.1$ they get the values $\alpha_0 = 0.0895$, $\alpha_1/\alpha = 0.318$ ($\alpha_0 = 0.212$, $\alpha_1/\alpha = 0.273$) under H_1 (H_0).

————— Insert Table 1 about here —————

3.2 The procedure without early acceptance

The calculations for the case without early stopping are very similar. For all choices $\alpha = 0.01, 0.025, 0.05$ the optimum is $\alpha_1 = c_\alpha$ (non-stochastic curtailment). This is plausible since a higher α_1 ($> c_\alpha$) leads to a higher power at the first stage and hence a higher overall power by systematically applying the sample size adaptation at the second stage. The optimal n_1 depends only on α and β . E.g. for $\alpha = 0.025$ and $\beta = 0.1$ (0.2) the optimal n_1 value is 42.2% (38.7%) of the sample size required for the fixed size sample z -test. Here under the null hypothesis the optimal n_1 is 0 as can easily be seen.

4 The influence of a priori misspecification of the variability

We now investigate the average sample size and the average power of the two procedures in the finite case for several parameter choices (for the derivations see Appendix B).

4.1 The procedure with early acceptance

We consider two choices of the acceptance boundary. First, we use the optimal α_0 that minimizes the sample size under the alternative. Second, we choose $\alpha_0 = 0.5$, i.e. we accept only if after the first stage there is no positive trend as suggested e.g. by Wieand and Therneau (1987) for the comparison of two independent binomial distributions. Table 2 gives the average power and average sample size for several parameter choices and an optimal n_1 of 36 ($\beta = 0.2$) and 44 ($\beta = 0.1$) resulting from an effect size of 1/2 of the standard deviation. We see that if the initial standard deviation estimate $\hat{\sigma}_0$ is too low the power

of the standard t -test collapses (last column). Note that n_{fix} here is calculated for the fixed size sample test from the corresponding t -distribution. As long as the variance was correctly chosen a priori the planned power is nearly met by the adaptive procedure. One reason for the small deviation may be the normal approximation in the sample size formulas. In the case where σ is overestimated the average power is exceeding the targeted power. However, if the variance was underestimated in the planning phase the procedure accepts H_0 too often after the first stage such that the overall power is noticeably below the planned power. This applies for the (small) optimal α_0 but – less severely – also for the case $\alpha_0 = 0.5$. However, in both cases the power is still higher than for the fixed size test.

To investigate robustness of the optimal design in the non-asymptotic case we looked at the results for small sample sizes as e.g. $n_1=9$ arising from an effect size of one standard deviation and $\alpha = 0.025$, $\beta = 0.025$. If the correct a priori variance is used in the planning phase applying the asymptotically optimal design we found a loss of power of less than two percentage points compared to Table 2 accompanied with a slightly lower sample size under the null hypothesis.

————— Insert Table 2 about here —————

4.2 The procedure without early acceptance

Table 3 shows the statistical properties in this situation. Now, the product test with sample size reassessment always exceeds the targeted power. The excess in power for the situation where the exact parameter values are used in the planning phase is due to early rejection. In the cases where the initial estimate $\hat{\sigma}_0$ is correct or too low, the average sample size under the alternative is close to n_{fix} . In the

case where σ is overestimated as in the previous procedure the average sample size is larger than the one of the t -test based on correct a priori estimates.

Under the null hypothesis the average sample size becomes quite large in all three scenarios since in this case we often have to perform the sample size re-assessment with a large p_1 value because no early acceptance occurs. For practical applications these cases may be avoided by introducing a maximum total sample size. This as mentioned above has no impact on the type I error probability.

————— Insert Table 3 about here —————

5 Midtrial redesign of the decision boundaries

As demonstrated above for the procedure with early acceptance, if we a priori underestimate the variance and choose a too small n_1 the power decreases dramatically since then the α_0 is too small for the chosen sample size. This problem can be addressed by adapting α_0 after the first stage, using the estimated variance. If α_0 is adapted additionally either (a) α_1 or (b) α_2 have to be adapted according to equation (2). Then asymptotically for $\Delta \rightarrow 0$ such that $n_1 \rightarrow \infty$ we have $\lim_{n_1 \rightarrow \infty} \hat{\sigma}_1 = \sigma$ almost surely and thus, the procedure asymptotically still meets the α level.

The adaptation of α_0 to achieve the asymptotic power $1-\beta$ is straightforward using relation (6). Let $\hat{\xi} = \frac{\Delta\sqrt{n_1}}{\sqrt{2}\hat{\sigma}_1}$. We start with the simple case (b) where we adapt α_0 and α_2 . Here we choose α_0 to satisfy

$$\frac{1-\beta}{\beta} \Phi_{(\hat{\xi},1)}(z_{1-\alpha_0}) = 1 - \Phi_{(\hat{\xi},1)}(z_{1-\alpha_1}). \quad (8)$$

Then we determine α_2 by (2) which gives

$$\alpha_2(\alpha_0, \alpha_1) = 1 - X_4^2(-2 \log[(\alpha - \alpha_1)/\{\log(\alpha_0) - \log(\alpha_1)\}]), \quad (9)$$

where X_4^2 denotes the distribution function of the χ_4^2 distribution. The early rejection boundary α_1 is not changed.

In case (a) we fix $\alpha_2 = \alpha$ but modify α_0 and α_1 such that (2) and (8) are satisfied. To this end we replace in (8) α_0 by $\alpha_0(\alpha_1)$ defined in (3) and then solve for α_1 . Note that a decrease of α_0 leads to an increase of α_1 . This happens if the variance is smaller than expected and hence typically if the p -value for the first stage is small. Hence, in this case the early rejection boundary and thus, the probability of an early rejection is raised. This seems to be a rather provocative property from a practical viewpoint. In case (b) where α_2 is modified, a decrease in α_0 leads to an increase in α_2 . Starting with $\alpha_2 = \alpha$, this leads to a final test with a level greater than α . Therefore for both cases (a) and (b) we study only a modified procedure which allows α_0 only to be raised (if the observed variance is larger than $\hat{\sigma}_0$). Thus, an overestimation of the variance should lead to a power exceeding $1 - \beta$.

In both cases for $\Delta \rightarrow 0$ such that $\hat{\sigma}_1 \rightarrow \sigma$ almost surely the test asymptotically controls the level α and (because of the one sided adaptation of α_0) exceeds the planned power $1 - \beta$. However, what happens in the finite case? We compute the probability that H_0 is rejected under the null and the alternative hypotheses. Table 4 gives the results for several parameter choices. The formula for the average power is given in Appendix B. For both cases (a) and (b) the power does not drop noticeably and the type I error probability is hardly inflated. Adapting α_2 instead of α_1 makes the procedure even less anti-conservative. The average sample sizes under H_0 as well as H_1 are quite moderate.

It may be suspected that applying very low sample sizes could increase the type I error probability substantially. Performing the calculations for $n_1 = 9$

(a priori and actual effect size of one standard deviation, $\alpha = 0.025, \beta = 0.1$) and case (b) increased the type one error probability only from 0.0251 to 0.0253 accompanied with a practically unchanged power but a slight increase of the relative average sample size. Note that the type I error probabilities for the mid-trial design modification hold independently of the sample size reassessment rule applied: under H_0 the distribution of p_2 is independent from this rule.

————— Insert Table 4 about here —————

6 Application

The application of the procedures is straightforward. When allowing for early acceptance optimal α_0 and α_1 -values (leading to an optimal n_1 by solving (8)) have to be determined first. For the procedure not allowing early acceptance the optimal α_1 is c_α such that only the optimal n_1 has to be determined. All the calculations can be performed with a Mathematica notebook available under <http://www.mstat.univie.ac.at/reassess/>.

6.1 An Example

To demonstrate the application of the method we sketch a theoretic example from urology. Various trials have been run to treat patients with voiding disorders associated with benign prostatic hyperplasia using the WHO International Prostatic Symptoms Score (IPSS, $0 \leq IPSS \leq 35$) as the primary efficacy variable.

An active treatment is compared to placebo. The pre-post difference after 6 months has been chosen as primary variable. A standard deviation of 5 units is assumed in the planning phase for the pre-post difference (based on previous

trials). We assume the relevant advantage over placebo to be 2 points on the IPSS scale. With $\alpha = 0.025$, $1 - \beta = 0.9$ the optimal design for a trial including early acceptance according to Section 3.1 would be $\alpha_0 = 0.206$, $\alpha_1 = 0.0150$, and $n_1 = 0.524 \cdot n_{\text{fix}} = 70$, where $n_{\text{fix}} = 133$. In the interim analysis $\hat{\sigma}_1 = 6.1$ and $p_1 = 0.21$ based on the one sided two sample t -test has been observed.

The redesign of the trial has been defined in advance to be performed by adjusting α_0 and α_2 . Therefore we get $\alpha'_0 = 0.402$ and $\alpha'_2 = 0.0207$ ($c_{\alpha_2} = 0.00304$). Using $p_1 = 0.21$ we get an n_2 of 224. Thus, the overall sample size is 294 per group which is about 50% more than needed for the fixed sample t -test planned with $\hat{\sigma}_0 = 6.1$. Note that a p -value of 0.21 or larger has a probability under the alternative (if 6.1 were the true standard deviation) of only 0.13 so that such extreme sample sizes are rather rare. Under the same alternative the probability of an early rejection is as large as 41% so that with a high probability a positive decision would be achieved with the rather small sample already at the first stage. Looking at the situation under the null hypothesis the procedure accepts early with a probability of about 0.6 (if integrating over all redesign scenarios with a true $\sigma = 6.1$). To be complete, early rejection will occur with an additional probability of 0.015 (the early rejection boundary is not adapted).

7 Concluding Remarks

We have reduced the flexibility of general adaptive two stage designs to sample size reassessment based on conditional power arguments.

Note first that by construction of the adaptive procedure we get sequential level α tests even if it is decided in the interim analysis to choose the sample size based on information beyond the estimated variability, e.g. the estimated

effect size. This extra flexibility together with the option of early stopping make comparisons with other methods of sample size reassessment difficult and questionable.

First we looked at “optimal” designs including an early acceptance decision. Given that all assumptions in the planning phase are correct we can construct decision rules that lead to an overall power close to the intended one when applying systematic sample size reassessment. When the variability has been underestimated a priori the design optimal under the a priori specification may, under the alternative, lead to a considerable loss of power due to frequent early acceptance of the null hypothesis. This does not occur when early acceptance is not included. Then as in procedures without early stopping options systematic sample size reassessment leads to satisfactory power values for the price of larger average sample sizes under the null hypothesis.

A more sophisticated way to avoid a loss of power in case of an a priori underestimation of the variance is to use the observed variability from the first stage to perform a mid-trial redesign of the acceptance limit. In particular, if the observed variability is larger than the a priori estimate we propose to make early acceptance less probable. Consequently, in order to preserve the overall level α the rejection decision has to be made more difficult too. It can be shown that such a procedure with systematic sample size reassessment does not noticeably inflate the type I error probability but preserves the intended power. Overall this result is not completely unexpected. First, the estimate of the variability for the normal distribution is stochastically independent of the estimated difference in the means. Second, the changes in acceptance limits are paid back by changes in the rejection limits. Both decisions are made more difficult in case of a priori

underestimation of the variability.

These findings are again an indication that the mid-trial variance estimate may be used in various ways without noticeably compromising the statistical properties of decision procedures on the mean of a normal distribution.

Acknowledgement We thank Meinhard Kieser, the editor and referees for constructive comments.

A Appendix: Minimizing the sample size

We discuss here the minimization of sample size for the case with early acceptance (see Section 3.1). The formulas for the case without early acceptance are very similar. In the following we assume that σ is known. Thus, we use the normal distribution throughout and insert σ in (4) and (7).

Let Δ_R denote the *real* difference between the means of the two groups. The distribution of p_1 is then given by $f(p) = \frac{\varphi(\frac{\Delta_R \sqrt{n_1}}{\sqrt{2}\sigma}, 1)^{(z_1 - p_1)}}{\varphi_{(0,1)}(z_1 - p_1)}$, where $\varphi_{(\mu, \sigma)}$ denotes the probability density function of the normal distribution with parameters μ and σ . If n_1 is chosen by (7) then under the alternative hypothesis if $\Delta_R = \Delta$ the density of p_1 can be written as $f_{H_1}(p_1) = \frac{\varphi_{\{\xi(\alpha_0, \alpha_1), 1\}}(z_1 - p_1)}{\varphi_{(0,1)}(z_1 - p_1)}$, where $\xi(\cdot)$ is defined by (6). Note that f_{H_1} is independent of σ and Δ_R . The average total sample size under the alternative in each group is given by

$$\begin{aligned} \bar{n}(\alpha_0, \alpha_1) &= n_1(\alpha_0, \alpha_1) + \int_{\alpha_1}^{\alpha_0} n_2\{p_1, \sigma, \alpha_2(\alpha_0, \alpha_1)\} f_{H_1}(p_1) dp_1 \\ &= \frac{2\sigma^2}{\Delta^2} \left\{ \xi(\alpha_0, \alpha_1)^2 + \int_{\alpha_1}^{\alpha_0} (z_\beta + z_{c_\alpha/p_1})^2 f_{H_1}(p_1) dp_1 \right\}, \end{aligned} \quad (10)$$

where $n_1(\cdot)$ is given by (7) and $\alpha_2(\cdot)$ by (9). Note that the average sample size under H_1 only depends on α, β, α_0 and α_1 . It can be easily shown that the

same holds under H_0 . We now minimize the average sample size numerically under the constraint $\alpha_2 = \alpha$ such that for given α_1 the boundary α_0 is given by (3). This leads to a one dimensional minimization which we performed with the Mathematica procedure FindMinimum.

B Appendix: Computation of the average power and sample size for unknown variance

Let Δ and $\hat{\sigma}_0^2$ be the assumed difference and variance in the planning phase and Δ_R and σ^2 the true difference and variance. We give here a general form of the formulas that apply for the cases without and with mid-trial design modifications (a) and (b) in Section 5. For all cases we denote the rejection resp. acceptance boundaries by $\alpha'_1(\hat{\sigma}_1^2), \alpha'_0(\hat{\sigma}_1^2), \alpha'_2(\hat{\sigma}_1^2)$. The probability that H_0 is rejected after the first stage is given by $\pi_1 = P(p_1 < \alpha'_1(\hat{\sigma}_1))$. For method (a) p_1 and $\alpha'_1(\hat{\sigma}_1^2)$ are not independent. We have (for all methods)

$$\begin{aligned} \pi_1 &= P\left(\frac{Z}{\sqrt{\frac{U}{\nu}}} > t_{\nu, 1-\alpha'_1(U\sigma^2/\nu)}\right) \\ &= P\left(Z > t_{\nu, 1-\alpha'_1(U\sigma^2/\nu)} \sqrt{\frac{U}{\nu}}\right) \\ &= 1 - \int_0^\infty \Phi_{\left(\frac{\Delta_R \sqrt{n_1}}{\sqrt{2}\sigma}, 1\right)}\left(t_{\nu, 1-\alpha'_1(u\sigma^2/\nu)} \sqrt{\frac{u}{\nu}}\right) f_{\chi_\nu^2}(u) du. \end{aligned}$$

where $\nu = 2n_1 - 2$, Z is $N\left(\frac{\Delta_R \sqrt{n_1}}{\sqrt{2}\sigma}, 1\right)$ distributed, U is χ_ν^2 distributed, $t_{\nu, \gamma}$ denotes the γ -quantile of the central t -distribution with ν degrees of freedom, $f_{\chi_\nu^2}$ is the density function of the χ^2 distribution with ν degrees of freedom, and $\Phi_{(\mu, \sigma)}$ denotes the cumulative distribution function of the normal distribution. Similarly, the probability that we stop after the first stage accepting H_0 is given by $\rho_1(n_1) = P\left(Z < t_{\nu, 1-\alpha'_0(U\sigma^2/\nu)} \sqrt{\frac{U}{\nu}}\right) = \int_0^\infty \Phi_{\left(\frac{\Delta_R \sqrt{n_1}}{\sqrt{2}\sigma}, 1\right)}\left(t_{\nu, 1-\alpha'_0(u\sigma^2/\nu)} \sqrt{\frac{u}{\nu}}\right) f_{\chi_\nu^2}(u) du$.

Hence, the probability of rejecting H_0 conditional on the event that we stop in the first stage is given by $\frac{\pi_1(n_1)}{\pi_1(n_1)+\rho_1(n_1)}$. If we reach the second stage the conditional probability to reject is $\pi_2(n_2, p_1, \alpha_2) = 1 - G\left(2n_2 - 2, \frac{\Delta_R \sqrt{n_2}}{\sqrt{2}\sigma}\right) \left(t_{2n_2 - 2, 1 - c_{\alpha_2}/p_1}\right)$, where $G_{(\nu, \xi)}$ denotes the cumulative distribution function of the t -distribution with ν degrees of freedom and non-centrality parameter ξ . The probability that the second stage is reached **and** H_0 is rejected is given by

$$\bar{\pi}_2 = \int_0^\infty \int_{t_{\nu, 1 - \alpha'_0(u\sigma^2/\nu)} \sqrt{\frac{u}{\nu}}}^{t_{\nu, 1 - \alpha'_1(u\sigma^2/\nu)} \sqrt{\frac{u}{\nu}}} \pi_2 \left[n_2 \left\{ p_1(z, u), \sqrt{\frac{u\sigma^2}{\nu}}, \alpha'_2\left(\frac{u\sigma^2}{\nu}\right) \right\}, p_1(z, u), \alpha'_2\left(\frac{u\sigma^2}{\nu}\right) \right] * f_{\chi_\nu^2}(u) \phi_{\left(\frac{\Delta_R \sqrt{n_1}}{\sqrt{2}\sigma}, 1\right)} dz du, \text{ where } p_1(z, u) = 1 - G_{\nu, \frac{\Delta_R \sqrt{n_1}}{\sqrt{2}\sigma}}\left(\frac{z}{u/\nu}\right), \text{ and } n_2(\cdot) \text{ is given by (4). The average power is then given by } \bar{\pi} = \pi_1 + \bar{\pi}_2. \text{ The average sample size is given by } n_1 + \int_0^\infty \int_{t_{\nu, 1 - \alpha'_0(u\sigma^2/\nu)} \sqrt{\frac{u}{\nu}}}^{t_{\nu, 1 - \alpha'_1(u\sigma^2/\nu)} \sqrt{\frac{u}{\nu}}} n_2 \left\{ p_1(z, u), \sqrt{\frac{u\sigma^2}{\nu}}, \alpha'_2\left(\frac{u\sigma^2}{\nu}\right) \right\} * f_{\chi_\nu^2}(u) \phi_{\left(\frac{\Delta_R \sqrt{n_1}}{\sqrt{2}\sigma}, 1\right)} dz du.$$

References

- Bauer, M., Bauer, P., and Budde, M. (1998). A simulation program for adaptive two stage designs. *Computational Statistics & Data Analysis* **26**, 351–371.
- Bauer, P. and Köhne, K. (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.
- Bauer, P. and Röhmel, J. (1995). An adaptive method for establishing a dose response relationship. *Statistics in Medicine* **14**, 1595–1607.
- Birkett, M. A. and Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine* **13**, 2455–2463.
- Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine* **14**, 1933–1940.

- Case, L. D., Morgan, T. M., and Davis, C. E. (1987). Optimal restricted two-stage designs. *Controlled Clinical Trials* **8**, 146–156.
- Coffey, C. S. and Muller, K. E. (1999). Exact test size and power of a gaussian error linear model for an internal pilot study. *Statistics in Medicine* **18**, 1199–1214.
- Cui, L., Hung, H. M. J., and Wang, S. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 321–324.
- Kieser, M. and Friede, T. (1999). Re-calculating the sample size of clinical trials in internal pilot studies with control of the type I error rate. *Statistics in Medicine* **19**, 901–911.
- Kieser, M. and Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal* **39**, 943–949.
- Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–1290.
- Lipták, T. (1958). On the combination of independent tests. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* **3**, 1971–1977.
- Mosteller, F. and Bush, R. (1954). Selected quantitative techniques. In Lindzey, G., editor, *Handbook of Social Psychology* pages 289–334. Addison-Wesley Cambridge, MA.
- Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal* **41**, 689–696.

- Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.
- Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190 – 197.
- Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243–258.
- Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics* **54**, 696–705.
- Wassmer, G. (1999). *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien*. Köln Verlag Alexander Mönch.
- Wieand, S. and Therneau, T. (1987). A two-stage design for randomized trials with binary outcomes. *Controlled Clinical Trials* **8**, 20–28.
- Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65–72.
- Wittes, J. T., Schabenberger, O., Zucker, D. M., Brittain, E., and Proschan, M. (1999). Internal pilot studies I: Type I error rate of the naive t-test. *Statistics in Medicine* **18**, 3481–3491.
- Zucker, D. M., Wittes, J. T., Schabenberger, O., and Brittain, E. (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine* **18**, 3493–3509.

β	0.1			0.2		
α	α_0	$\frac{\alpha_1}{\alpha}$	$\frac{n_1}{n_{\text{fix}}}$	α_0	$\frac{\alpha_1}{\alpha}$	$\frac{n_1}{n_{\text{fix}}}$
0.01	0.132	0.594	0.542	0.106	0.630	0.588
0.025	0.206	0.601	0.524	0.171	0.639	0.570
0.05	0.284	0.612	0.508	0.241	0.652	0.554

Tab. 1: Optimal α_0 , α_1 , and the corresponding n_1 as fraction of n_{fix} .

α	β	σ	α_0	ρ_1	π_1	$\bar{\pi}$	$\frac{\bar{n}_\Delta}{n_{\text{fix}}}$	$\frac{\bar{n}_0}{n_{\text{fix}}}$	π_{fix_0}
0.025	0.1	1	0.206	< 0.001	0.993	0.999	2.01	2.16	1
			0.5	< 0.001	0.890	0.989	1.27	1.71	
		2	0.206	0.0640	0.558	0.897	0.786	0.681	0.9
			0.5	0.0343	0.292	0.894	0.849	0.833	
		3	0.206	0.229	0.266	0.718	0.621	0.398	0.58
			0.5	0.112	0.128	0.805	0.828	0.667	
	0.2	1	0.171	< 0.001	0.978	0.995	2.15	2.26	1
			0.5	< 0.001	0.765	0.952	1.32	1.69	
		2	0.171	0.123	0.474	0.795	0.831	0.683	0.8
			0.5	0.0573	0.213	0.790	0.910	0.839	
		3	0.171	0.324	0.224	0.583	0.579	0.377	0.463
			0.5	0.146	0.0961	0.695	0.844	0.676	

Tab. 2: Average power and sample sizes for the procedure with early acceptance.

The results are given for the optimal α_0 , for $\alpha_0 = 0.5$, and the a priori standard deviation estimate $\hat{\sigma}_0 = 2$. ρ_1 denotes the probability of early acceptance under H_1 , σ the actual standard deviation, π_1 the probability to reject at the first stage under H_1 , $\bar{\pi}$ the average overall power, $\bar{n}_\Delta/n_{\text{fix}}$ and \bar{n}_0/n_{fix} the average sample size under H_0 and H_1 as fraction of the sample size of the one stage t -test planned with the actual variance, and π_{fix_0} the power of the one stage t -test planned with $\hat{\sigma}_0 = 2$.

α	β	σ	π_1	$\bar{\pi}$	$\frac{\bar{n}_\Delta}{n_{\text{fix}}}$	$\frac{\bar{n}_0}{n_{\text{fix}}}$	$\pi_{\text{fix}0}$
0.025	0.1	1	0.894	0.992	1.51	2.65	1
		2	0.239	0.922	0.895	1.60	0.9
		3	0.0866	0.903	0.966	1.40	0.580
	0.2	1	0.800	0.972	1.66	2.78	1
		2	0.181	0.837	0.991	1.67	0.8
		3	0.0671	0.809	1.03	1.46	0.463

Tab. 3: Average power and average sample sizes for the procedure without early acceptance choosing $\alpha_1 = c_\alpha$ and n_1 optimal for $\Delta = 1$ and the a-priori standard deviation $\hat{\sigma}_0 = 2$. The columns are defined as in Table 2.

α	β	σ	ad.	α_{eff}	ρ_1	π_1	$\bar{\pi}$	$\frac{\bar{n}_\Delta}{n_{\text{fix}}}$	$\frac{\bar{n}_0}{n_{\text{fix}}}$	$\pi_{\text{fix}0}$
0.025	0.1	1	α_1	0.0250	< 0.001	0.993	0.999	2.00	2.16	1
			α_2	0.0250	< 0.001	0.993	0.999	2.00	2.16	
		2	α_1	0.0252	0.0527	0.549	0.907	0.804	0.723	0.9
			α_2	0.0251	0.0534	0.557	0.907	0.806	0.723	
		3	α_1	0.0252	0.026	0.197	0.894	0.892	0.998	0.58
			α_2	0.0250	0.0334	0.265	0.894	0.931	0.989	
	0.2	1	α_1	0.0250	< 0.001	0.978	0.995	2.15	2.26	1
			α_2	0.0250	< 0.001	0.978	0.995	2.15	2.26	
		2	α_1	0.0252	0.102	0.465	0.812	0.862	0.726	0.8
			α_2	0.0251	0.104	0.474	0.812	0.867	0.724	
		3	α_1	0.0253	0.0475	0.163	0.793	0.941	0.977	0.463
			α_2	0.0251	0.0626	0.224	0.794	0.998	0.96	

Tab. 4: Average power and sample sizes for the procedure with early acceptance and redesign of the decision boundaries. The columns are defined as in Table 3 and 2. The column 'ad.' describes the two scenarios of increasing α_0 by adjusting α_1 ($\alpha_2 = \alpha$) or α_2 (α_1 unchanged) and the column α_{eff} denotes the effective type I error probability.