Published in final edited form as:

Stat Appl Genet Mol Biol. 2015 November ; 14(5): 429–442. doi:10.1515/sagmb-2014-0025.

Sample size reassessment for a two-stage design controlling the false discovery rate

Sonja Zehetmayer*, Alexandra C. Graf, and Martin Posch

Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, Vienna, Austria

Abstract

Sample size calculations for gene expression microarray and NGS-RNA-Seq experiments are challenging because the overall power depends on unknown quantities as the proportion of true null hypotheses and the distribution of the effect sizes under the alternative. We propose a two-stage design with an adaptive interim analysis where these quantities are estimated from the interim data. The second stage sample size is chosen based on these estimates to achieve a specific overall power. The proposed procedure controls the power in all considered scenarios except for very low first stage sample sizes. The false discovery rate (FDR) is controlled despite of the data dependent choice of sample size. The two-stage design can be a useful tool to determine the sample size of high-dimensional studies if in the planning phase there is high uncertainty regarding the expected effect sizes and variability.

Keywords

adaptive design; false discovery rate; high-dimensional data; two-stage design

1 Introduction

The analysis of gene expression microarray and NGS-RNA-Seq experiments typically involves a large number of hypothesis tests. Sample size calculations for such studies are challenging because the overall power depends on the unknown proportion of null hypotheses and the unknown effect size distribution for the hypotheses where the alternative holds (Jung, 2005; Pawitan et al., 2005; Tibshirani, 2006). One approach to address this problem are sequential designs where frequent interim analyses are performed and the experiment is stopped as soon as the estimated power exceeds a certain threshold (Posch et al., 2009). These designs, however, can require a large number of interim analyses which may not be practical and computationally expensive.

In this paper we propose a sample size reassessment procedure for studies where a large number of hypotheses is tested controlling the false discovery rate (FDR). In an interim

*Corresponding author: Sonja Zehetmayer, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, Vienna, Austria, sonja.zehetmayer@meduniwien.ac.at.

Supplemental Material: The online version of this article (DOI: 10.1515/sagmb-2014-0025) offers supplementary material, available to authorized users.

analysis the distribution of the effect sizes under the alternative is estimated with an empirical Bayes mixture model (Muralidharan, 2010) or a kernel deconvolution density estimator (van Iterson et al., 2013). Based on the estimated effect size distribution, the sample size of the second stage is determined such that the predicted overall power at the end of the experiment (defined as the expected proportion of rejected hypotheses among all hypotheses for which the alternative holds) exceeds a pre-specified threshold. If there is a large fraction of alternative hypotheses with very small effect sizes, this overall power may not be useful because alternative hypotheses with very low effect sizes are often of little interest. Therefore, we alternatively propose the control of the contingent power, defined as the proportion of rejected hypotheses among the alternative hypotheses with an effect size that exceeds a threshold (Matsui and Noma, 2011).

Typically approaches to estimate the power (or equivalently the multiple Type II error) in high-dimensional experiments are based on estimators of the fraction of true null hypotheses (DeLongchamp et al., 2004; Norris and Kahn, 2006; Zehetmayer and Posch, 2010). While these procedures allow to estimate the post-hoc power after all observations have been collected, they do not allow for a prospective power calculation because the latter requires an estimation of the distribution of effect sizes.

The adaptive sample size reassessment procedure for the high-dimensional testing problem considered here is based on standard elementary fixed sample test statistics and multiplicity adjustments to control the FDR. Under suitable distributional assumptions, similarly to the sequential testing procedures studied in Posch et al. (2009), the adaptive test procedure approximately controls the FDR in spite of the adaptation of the sample sizes. This is in contrast to sample size reassessment for tests of a single hypothesis that may lead to a severe Type I error rate inflation (Proschan and Hunsberger, 1995; Graf and Bauer, 2011) if not accounted for, e.g. by applying tests based on combination functions or conditional error rate based tests (Bauer and Köhne, 1994; Posch and Bauer, 1999; Müller and Schäfer, 2001; Brannath et al. 2002; Posch et al., 2003).

Below we introduce the sample size reassessment procedure and the concept of overall and contingent predictive power. In the results section we report a simulation study investigating the operation characteristics of the procedure and give a real data example. We close with a short discussion.

2 Methods

Consider an experiment to test m means μ_i , $i=1, \dots, m$, of independent normally distributed observations and unknown variances σ_i^2 . Each hypothesis $H_{0i}:\mu=0$ versus $H_{0i}:\mu \neq 0$, $i=1, \dots, m$, is tested with a one-sample t-test with p -value p_i . The FDR (defined as the expected fraction of false rejections under all rejections) is controlled at level α with the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) which rejects all hypotheses H_{0i} with $p_i \leq d/\alpha$ where $d=\arg\max_i\{p(i) \leq \alpha/m\}$ with $p(i)$ denoting the ordered p -values. The BH procedure controls the FDR if the test statistics are independent or positively dependent (Benjamini and Yekutieli, 2001).

2.1 Sample size calculation for fixed sample experiments controlling the FDR

Sample size calculations for experiments with a large number of hypotheses allow to specify sample sizes that guarantee control of a multiple Type II error rate or multiple power measure at a pre-specified threshold. A frequently considered power concept is that of the overall power, defined as the expected proportion of alternative hypotheses that are rejected. As hypotheses with a very small effect size have little relevance in many settings, we consider in addition the contingent power Pow_Δ [compare also Matsui and Noma (2011)]. The contingent power denotes the expected proportion of rejected alternative hypotheses among the alternative hypotheses with a standardized effect size of at least Δ . Thus, denoting the actual (unknown) effect sizes by $\delta_i, i=1, \dots, m$ (where $\delta_i=0$ for all true null hypotheses H_0), the contingent power is a conditional power given by

$$\text{Pow}_\Delta = \frac{E(\#\{i | p_i \leq \gamma_{\text{BH}}, |\delta_i| \geq \Delta\})}{\#\{i | |\delta_i| \geq \Delta\}},$$

where γ_{BH} denotes the BH critical value. Setting $\Delta=0$, we obtain as special case the overall power Pow_0 .

The required sample size depends on the nominal FDR level and the actual effect sizes $\delta_i, i=1, \dots, m$. For the sample size calculation we assume that for a proportion π_0 of all hypotheses the null hypothesis holds (such that $\delta_i=0$) and the standardized effect sizes δ_i under the alternative are drawn from a distribution $g(\delta)$. Given the distribution $g(\delta)$ and π_0 an asymptotic expression for the power is given by

$$\text{Pow}_\Delta(n, \gamma^*) = \frac{\int_{\delta > \Delta} [1 - T_{n-1, \delta \sqrt{n}}(t_{1-\gamma^*, n-1})] [g(\delta) + g(-\delta)] d\delta}{\int_{\delta > \Delta} [g(\delta) + g(-\delta)] d\delta}, \quad (1)$$

where n denotes the sample size, T_{n-1} the cumulative distribution function of the t -distribution with $n-1$ degrees of freedom and non-centrality parameter $\delta \sqrt{n}$, $t_{1-\gamma^*, n-1}$, the $1-\gamma^*$ -quantile of the t -distribution with $(n-1)$ degrees of freedom, and the asymptotic critical value of the BH procedure γ^* solves in γ

$$\alpha = \frac{\gamma}{\pi_0 \gamma + (1 - \pi_0) \text{Pow}_0(n, \gamma)}. \quad (2)$$

Some comments: (i) For $\Delta=0$, the denominator of (1) is equal to 1 and the nominator is a weighted average over the power of the hypothesis tests over a distribution of effect sizes. For $\Delta > 0$ the average is taken over a truncated range of effect sizes and normalized accordingly by the proportion of alternatives with absolute effect size larger than Δ . (ii) The right hand side of (2) gives an asymptotic expression of the FDR for the BH procedure for large m (Genovese and Wasserman, 2002). As in Genovese and Wasserman (2002) or Zehetmayer et al. (2005) we exploit here the fact that for $\pi_0 < 1$ the critical value of the BH procedure converges for large m (and some conditions on the dependence structure of the

test statistics) almost surely to γ^* . Note that $\gamma^* = \gamma^*(n)$ is a function of n that also depends on $g(\delta)$ and π_0 . Thus, for a given sample size n , $\text{Pow}_0(n, \gamma^*)$ gives the overall power and $\text{Pow}(n, \gamma^*)$ the contingent power, where $\gamma^* > 0$. To compute the sample size required to achieve a certain target power CP^* the equation $\text{Pow}(n, \gamma^*) = \text{CP}^*$ is solved in n , for example, by a bisection algorithm.

2.2 Sample size reassessment

Assume that after a first stage with n_1 observations an interim analysis is performed. For each hypothesis i a first stage one-sample t-statistic $z_i, i=1, \dots, m$ is computed. Based on these asymptotically normally distributed z-scores we estimate the fraction π_0 of true null hypotheses as well as the distribution of standardized effect sizes $g(\delta)$ with the empirical Bayes mixture model approach by Muralidharan (2010). The z-scores z_i and effect sizes δ_i are assumed to be generated by the hierarchical sampling procedure $\delta \sim g(\delta)$ and $z_i | \delta \sim f_{\delta_i}(z)$ (Brown-Stein model). For the special case of a one-sample test of means of normally distributed observations we use the approximation $f_{\delta_i}(z) = \phi_{\delta_i \sqrt{n_1+1}}(z)$, where $\phi_{\mu, \sigma}$ denotes the normal density with mean μ and standard deviation σ . The distribution of the effect sizes is approximated by a mixture model with a point mass $\pi_0 \in [0, 1]$ at 0 (representing the null hypothesis) and a single normal distribution component to model the distribution of the effect sizes under the alternative. The normal component is given by $g(\delta) = \phi_0, \sigma(\delta)$ such that larger σ correspond to settings with larger effect sizes under the alternative. The parameters π_0 and σ are estimated based on a marginal maximum likelihood and the EM algorithm implemented in the R-package mixfdr (Muralidharan, 2012). The resulting estimates $\hat{\pi}_0, \hat{g}$ are then plugged into the power formula (1). If the power estimate meets the threshold already in the interim analysis, the trial is stopped and analyzed at the interim analysis. Otherwise, the second stage sample size is determined as

$$n_2 = \min(\min\{n | \text{Pow}_\Delta(n, \gamma^*) \geq \text{CP}^*\}, n_{\max}) - n_1,$$

where $n_{\max}$ is a fixed maximum overall sample size. Note that when estimating the post-hoc power for the analysis of the interim data only (assuming no additional observations are collected) the actual BH critical value γ_{BH} in (1) is used instead of the approximation γ^* .

2.3 Extension to two sample tests

To apply the sample size reassessment procedure for the two-sample t-test, let r denote the allocation fraction such that $n_A = rm$ and $n_B = (1-r)n$ denote the per group sample sizes and n the overall sample size. To obtain the correct power formula for the two sample case we replace $T_{n-1, \delta \sqrt{n}}(\gamma)$ by $T_{n-2, \delta \sqrt{nr(1-r)}}(\gamma)$ in (1). For the estimation of the mixture distribution, the density of the z-scores conditional on $\delta = \delta_i$ is approximated by $f_{\delta_i}(z) = \phi_{\delta_i \sqrt{n_1 r(1-r)+1}}(z)$, where n_1 denotes the overall first stage sample size.

2.4 Alternative approaches to estimate π_0 and the effect size distribution g

Besides the Bayesian mixture model estimator described above with a single mixture component to model the effect size distribution under the alternative, we consider two

alternative estimators of $g(\delta)$ for the numerical comparisons. First, a direct extension of the single component model where $g(\delta)$ is modeled with two normal densities whose means and variances are estimated. Second, we investigate the kernel deconvolution density estimator implemented in the R package SSPA by van Iterson et al. (2013). It can be applied for two- or multi-group comparisons where T-, F- or χ^2 -test statistics are applied. The latter, e.g. arise in RNA-seq experiments.

3 Results

3.1 Setup of the simulation study

The operation characteristics of the procedure are investigated in a simulation study for one- and two-sample t-tests. If not specified otherwise, sample size reassessment was performed to control the overall power ($\beta=0$) as well as the contingent power with threshold $\beta=0.1$ for the absolute effect size at 80%. The nominal FDR was set to $\alpha=0.05$. The reassessed second stage sample size was cut off at a maximum sample size of 400.

Simulations for two-sided one sample t-tests of $m=1000$ and $m=10,000$ hypotheses were performed, assuming that the effect sizes are drawn from a mixture distribution with a point mass at 0 and a normal mixture component centered at 0 with a standard deviation of 1. We refer to the latter scenario as the unimodal effect size distribution. To estimate the mixture model parameters π_0 and σ we applied the `mixModelFitter` function of the `mixfdr` R-package with a single mixture component (Muralidharan, 2012; R Core Team, 2013) to model the effect size distribution under the alternative. As the SSPA package does not cover one-sample tests the latter approach was considered only for the two-sample test scenarios.

We considered different dependence structures for the test statistics across hypotheses: Independence, a block correlation structure (also called clumpy correlation) or equicorrelation. Block correlation may be induced in microarray data for example by pathways of genes that are commonly regulated (Qiu et al., 2005). We assumed that the test statistics are correlated in blocks of 10 hypotheses. Within each block the data is equicorrelated with $\rho=0.5$ (Storey et al., 2004). Data from hypotheses of different blocks are assumed to be independent. Second we consider equi-correlation which can occur due to array effects in microarray analyses. For equi-correlation we assume that for all pairs of hypotheses a pairwise correlation of $\rho=0.5$ holds. For all correlation structures the alternatives are randomly distributed among the sequence of hypotheses and the same scenarios as for independent data are examined with $\sigma_i=1$, $i=1, \dots, m$. A number of 1000 simulation runs are performed per scenario for $\pi_0=0.8$ and $\pi_0=0.9$, 10,000 simulation runs for $\pi_0=0.99$.

In addition to simulations for independent test statistics as for the one-sample t-test, for the two-sample case we also investigated correlation structures of actual microarray data sets by simulating data from the lung disease microarray data set described in Section 4 following the lines of Qu et al. (2012). The simulated case and control group data were generated as follows: first, for both groups data arrays were sampled without replacement from the group of patients with interstitial lung disease (as it was the largest group in the data set with 254 observations). Then for the arrays in the case group effect sizes $\theta_i \hat{\sigma}_i$ were added to the

expression value of the i -th gene, where $\hat{\sigma}_i$ is the observed standard deviation of the values of gene i . θ_i was drawn from a mixture distribution with weight π_0 on 0 and weight $(1-\pi_0)$ on an $N(0, 1.5)$ distribution for the unimodal case or a bitriangular distribution [with parameters as recommended in (Langaas et al., 2005)] to model settings where the effect sizes follow a bimodal distribution under the alternative. We refer to the two scenarios as unimodal and bimodal effect size distributions. From the total number of $m=15,261$ genes the alternative hypotheses were randomly selected from the subset of genes with “biological process” ontology terms with a term size larger than 50 in the Gene Ontology Project Consortium (2000) data base. The maximum sample size in this simulation was determined by the sample size of the data set and set to 127. In the simulations with unimodal distributed effect sizes, the effect size distribution was estimated with the `mixModelFitter` function of the `mixfdr` R-package with a mixture model with point mass at 0 and a normal density mixture component centered at 0. For the bimodal case a mixture model with point mass at 0 and two normal density mixture components with estimated means and variances was used. In addition, we applied the SSPA approach to estimate both effect size distributions. In the simulations for the bimodal case we aimed to control the contingent power for $\alpha=0.5$ because the contingent power at $\alpha=0.1$ (the value used in the unimodal case) is very close to the overall power in this setting.

3.2 Results of the simulation study

Figure 1 shows the asymptotic power [computed from (2) and (1)] of a fixed sample test as function of sample size under a range of settings. The sample size to guarantee a certain overall or contingent power depends sensitively on the distribution of the effect sizes under the alternative. As expected, for the control of the contingent power lower sample sizes are needed to achieve a pre-specified power threshold. Tables 1 and 2 show the properties of the adaptive design with sample size reassessment for the unimodal setting of $m=1000$ and $m=10,000$ one-sample hypotheses tests with independent test statistics. Figure 2 shows the distribution of the sample size and proportion of rejected alternative hypotheses for the scenario $m=10,000$, $\sigma=1$.

The FDR is controlled in all considered scenarios. For $m=10,000$ and a first stage sample size of $n_1=50$ the actual power is within one percentage point of the target value of 80% in all scenarios except those where the first stage sample size already leads to a larger power. In the latter scenarios the expected sample size is equal (up to rounding) to the first stage sample size. For lower first stage sample sizes or a lower number of hypotheses the actually achieved overall and contingent power values range between 74%–84%. With first stage sample sizes of $n_1=30$ or $n_1=50$, also for the case of $m=1000$ hypotheses and $\pi_0=0.8, 0.9$ the actual power is within one percentage point of the target power (again with the exception of the cases where the target power is exceeded with the first stage sample). Only for $\pi_0=0.99$ the sample size reassessment procedure tends to too large sample sizes. The variability of the sample size is largest if π_0 is close to one for $m=1000$ but decreases for $m=10,000$ as well as lower values of π_0 . We additionally performed simulations under the global null hypothesis ($\pi_0=1$) and found that the procedure still controls the FDR well (all observed FDRs < 0.05) and in most runs the sample size is reassessed to $n_{\max}$.

For correlated test statistics the achieved power values are still close to the target power, however, the variability of the sample sizes increases compared to the case of independent test statistics (See Tables S1 and S2). Note that also for all investigated scenarios with correlated test statistics the FDR is controlled.

Table 3 shows the performance of the procedures based on the Bayesian mixture model (mixfdr) and the SSPA package for unimodal distributed effect sizes and independent test statistics in the two sample case. Note that we set the standard deviation of the effect size distribution to 1.5 (compared to 1 for the one sample case) such that larger effect sizes become more likely and the power of 80% can be reached with similar sample sizes as in the one-sample case. Similar to the one sample case, the mixfdr procedure provides a good control of the overall and the contingent power in all considered scenarios. The SSPA procedure in contrast gives in most scenarios highly variable estimates of the effect size distribution resulting in a large standard deviation of the respective sample sizes. Only for a large number of hypotheses ($m=10,000$) and π_0 not too close to 1 the procedure showed a reliable control of the contingent power.

For the simulations based on the microarray data set Table 4 shows the results for unimodal and bimodal distributed effect sizes. In the unimodal scenario the mixfdr procedure controls the overall power well with deviations up to 4 percentage points to the target power of 80%. As above, the SSPA method leads to more variable sample size and larger deviations for the target power, especially for π_0 close to 1. This is in concordance with the findings of van Iterson et al. (2013) who also observed that the estimate becomes unstable in these scenarios. Also the contingent power is well controlled by both procedures. Again SSPA leads to a more variable sample size (with the exception of the cases $\pi_0=0.99$, $n_1=30$, 50) where the SSPA procedure always chooses the maximum sample size 127.

For the bimodal effect size distribution, the mixfdr procedure (based on a two component mixture distribution of the effect sizes under the alternative) achieves still a good control of the overall power for $n_1=30$. For $n_1=10$ (as well as $n_1=50$ where the first stage power already exceeds 80%) we observe larger deviations from the target power. In addition, for most scenarios the variability of sample sizes is larger than in the unimodal case. The SSPA procedure leads to more variable sample sizes than the mixfdr procedure and tends to overestimate the required sample size. The mixfdr procedure severely underestimates the sample size required to control the contingent power, especially for $n_1=10$. In contrast the SSPA procedure overestimates the required sample size.

Table 5 shows results for a non-symmetric bimodal distribution of effect sizes under the alternative where the positive effect sizes dominate. The results are similar to the symmetric case. Table 6 shows the performance of the mixfdr procedure if the effect size distribution is misspecified, i.e. where the actual effect size distribution under the alternative is unimodal but modeled by a mixture of two normal distributions and vice versa. In the former case mixfdr gives too low sample sizes, because effect sizes are overestimated. On the other hand, if the bimodal effect size distribution is fitted by a single normal density centered at 0, effect sizes are underestimated, and always the maximum sample size is chosen. This effect is alleviated for the contingent power. For comparison the results of the SSPA procedure in the

considered scenarios are given. Finally, Table 7 shows simulations for different choices of the threshold α in the contingent power to illustrate its impact on the sample size.

4 Application to a microarray gene expression study

We retrospectively applied the adaptive sample size reassessment procedure to the gene expression profiling study of chronic lung disease (Geo database, series number GSE47460: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47460>) of the Lung Genomics Research Consortium (Edgar et al., 2002). In this experiment, gene expression of 220 COPD patients, 254 patients with interstitial lung disease and 108 controls was measured with Agilent microarrays giving measurements for $m=15,261$ genes. The normalized log₂-based signal intensities available for download were further normalized by replacing expression values with a normal transformation of the ranks (see Efron, 2010). Furthermore, we filtered the data by removing all genes with a gene-wise interim variance below 0.1.

For illustrative purposes, we performed separate analyses for the three pairwise comparisons between the disease groups applying two-sided two-sample t-tests. Because the allocation ratios in the three groups [control, chronic obstructive pulmonary disease (COPD), interstitial lung disease] are roughly 1:2:2 we used the corresponding sample size ratios for the pairwise comparisons. As in the simulation study we considered three different first stage per-group sample sizes.

The second stage sample size was chosen to achieve an (estimated) overall or contingent (with $\alpha=0.1$) power of 0.8 for a specific comparison. At the end of the trial, the overall power was again estimated based on the final data sets. Sample sizes, the estimated parameters of the effect size distribution and estimated power values for the example are given in Table 8.

To control the overall power ($\alpha=0$), the reassessed sample size would have exceeded the available sample and the total sample was used (up to 4 observations that were dropped to obtain the desired sample size ratios). For all, but the comparison of Control to COPD the estimated power at the end of the trial is close to the nominal level of 80%. Only for the comparison COPD/control the estimated power is considerably lower than the nominal value, because there were not sufficient observations available to increase the sample size according to the sample size reassessment rule. For the control of the contingent power ($\alpha=0.1$) the reassessed sample sizes are smaller such that the available sample size suffices for all comparisons except for control/COPD. With the latter exception the estimated power at the end of the experiment is close to the nominal level.

The R-code to estimate the power for different sample sizes based on an interim data set is available under <http://statistics.msi.meduniwien.ac.at/index.php?page=pageszfmr>.

5 Discussion

In this manuscript we proposed an adaptive two-stage design to reassess the sample size based on an interim estimate of the distribution of effect sizes under the alternative. Besides controlling the overall power we considered adaptation rules that aim to control the

contingent power, i.e. the power to reject alternative hypotheses with an effect size exceeding a certain threshold δ . This threshold can either be chosen as a minimum relevant difference that arises from the biological context or one can compute the reassessed sample size for several δ to make an appropriate choice. The simulation study demonstrated that the procedure adapts the sample size reliably to achieve a pre-specified power or contingent power. However, a too small first stage leads to an unreliable sample size estimation, especially in the scenario of only $m=1000$ hypotheses. This is most pronounced in settings where π_0 is very close to 1 and the mixture model estimates π_0 less accurately. We also observed that the sample size reassessment procedure is rather sensitive to the model assumptions and a misspecification of the model for the effect size distribution may lead to incorrect sample size estimates. In contrast, applying more flexible models as the `mixfdr` method fitting a bimodal distribution for the effect sizes under the alternative or the SSPA method may lead to an “overfitting” of the effect size distribution and more variable estimates. In all considered scenarios the procedure consistently controls the FDR at the pre-specified level.

There are essentially two reasons why the classical fixed sample tests are valid when the proposed sample size reassessment rule, which assigns to all hypotheses the same second stage sample size, is applied to tests for a large number of hypotheses. First, for the fixed sample test the BH procedure asymptotically controls the false discovery proportion (FDP), defined as the fraction of erroneously rejected hypotheses among all rejected hypotheses and not only the FDR, its expected value. This holds for sufficiently independent test statistics if for a positive fraction of hypotheses the alternative holds and additional technical conditions apply, cf. Storey (2002). Then the FDR of the adaptive test can be bounded from above by the FDR of a sequential test that continues to the maximal sample size but where the analysis is retrospectively performed at the time point where the FDP is maximal. However, as the number of hypotheses increases, the FDPs at each time point converge almost surely to a constant not exceeding the FDR. Thus, the level of this sequential test (and consequently also the FDR of the test under sample size reassessment) is asymptotically equal to α (Posch et al., 2009). The second reason why the classical test controls the FDR for large m is that the variability of the second stage sample size decreases with increasing numbers of hypotheses.

Several generalizations of the procedure can be considered: (i) Instead of performing a single interim analysis the sample size may be reassessed several times. In such a setting, instead of collecting all the observations of the second stage sample, an additional interim analysis is performed to allow a further adjustment of the sample size based on a larger sample. It is expected that such a strategy will further reduce the variability in sample size. (ii) The sample size reassessment procedure allows for an early stopping of the experiment at the interim analysis: if the estimated interim power already exceeds the target power, no second stage sample is collected. Additionally, one can implement a futility rule: if the reassessed sample size would be too large, one might consider to stop the experiment early. Still, the theoretical results in Posch et al. (2009) suggest that the FDR will be controlled in this case as well. (iii) We formulated the procedure for one and two sample tests for the mean of normally distributed observations. In many testing scenarios (e.g. the comparison of rates), tests with asymptotically normally distributed test statistics can be defined. To adapt

the procedure for specific testing procedures in the computation of the power (1) the cumulative t-distribution function can be replaced by the normal distribution with appropriate non-centrality parameters. To extend the sample size reassessment procedure to count data arising from RNA-Seq experiments, one can, e.g. transform the RNA-Seq data such that t-statistics from linear models can be computed (e.g. via the VROOM function in the limma package in bioconductor [Smyth, 2005]). Alternatively, the SSPA method to estimate the effect size distribution power can be used, which can be applied to a large number of test statistics including χ^2 distributed test statistics arising in RNA-seq experiments. (iv) Instead of the mixture model approaches by Muralidharan (2010, 2012) or the kernel deconvolution density estimator by van Iterson et al. (2013) alternative estimators of the effect size distributions can be used to estimate the power at different sample sizes, see, e.g. Qu et al. (2012), Matsui and Noma (2011). (v) An interesting question is if the procedure can be extended to tests that control the familywise error rate (FWER), as for example Bonferroni tests. Graf et al. (2014) have demonstrated, that general sample size reassessment rules for tests of multiple hypotheses may lead to a severe inflation of the FWER. However, as Posch et al. (2009) note, in settings where the reassessed sample size converges almost surely as the number of hypotheses m increases, asymptotically the FWER will not be affected by the sample size reassessment. For the sample size reassessment this is the case, if the estimate of the effect size distribution converges as $m \rightarrow \infty$. While Muralidharan (2010) does not prove the convergence of the estimator, the simulations suggest that the variability reduces with increasing m .

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: Austrian Science Fund, (Grant/Award Number: 'P23167').

References

- Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994; 50:1029–1041. [PubMed: 7786985]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*. 1995; 57:289–300.
- Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 2001; 29:1165–1188.
- Brannath W, Posch M, Bauer P. Recursive combination tests. *J. Am. Stat. Assoc.* 2002; 97:236–244.
- Consortium, T. G. O. Gene ontology: tool for the unification of biology. *Nat Genet.* 2000; 25(1):25–29. [PubMed: 10802651]
- Delongchamp RR, Bowyer J, Chen J, Kodell R. Multiple-testing strategy for analyzing cdna array data on gene expression. *Biometrics*. 2004; 60:774–782. [PubMed: 15339301]
- Edgar R, Domrachev M, Lash A. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nuc. Acids Res.* 2002; 30:207–210.
- Efron B. z-values and the accuracy of large-scale statistical estimates. *J. Am. Stat. Assoc.* 2010; 105:1042–1055. [PubMed: 21052523]
- Genovese C, Wasserman L. Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B*. 2002; 64:499–517.

- Graf AC, Bauer P. Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. *Stat. Med.* 2011; 30:1637–1647. [PubMed: 21495058]
- Graf AC, Bauer P, Glimm E, Koenig F. Maximum type 1 error rate inflation in multi-armed clinical trials with adaptive interim sample size modifications. *Biometrical J.* 2014; 56(4):614–630.
- Jung SH. Sample size for fdr-control in microarray data analysis. *Bioinformatics.* 2005; 21:3097–3104. [PubMed: 15845654]
- Langaas M, Lindqvist B, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. R. Stat. Soc. B.* 2005; 67:555–572.
- Matsui S, Noma H. Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. *Biometrics.* 2011; 67:1225–1235. [PubMed: 21627629]
- Müller H-H, Schäfer H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics.* 2001; 57:886–891. [PubMed: 11550941]
- Muralidharan O. An empirical bayes mixture method for effect size and false discovery rate estimation. *Annals of Applied Statistics.* 2010; 4:422–438.
- Muralidharan, O. mixfdr: computes false discovery rates and effect sizes using normal mixtures. r package version 1.02012. URL <http://CRAN.R-project.org/package=mixfdr>, r package version 1.0
- Norris AW, Kahn C. Analysis of gene expression in pathophysiological states: Balancing false discovery and false negative rates. *Proc. Natl Acad. Sci.* 2006; 103:649–653. [PubMed: 16407153]
- Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics.* 2005; 21:3017–3024. [PubMed: 15840707]
- Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical J.* 1999; 41:689–696.
- Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Stat. Med.* 2003; 23:953–969. [PubMed: 12627412]
- Posch M, Zehetmayer S, Bauer P. Hunting for significance with the false discovery rate. *J. Am. Stat. Assoc.* 2009; 104:836–840.
- Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics.* 1995; 51:1315–1324. [PubMed: 8589224]
- Qiu X, Brooks AI, Klebanov L, Yakovlev A. The effects of normalization of the correlation structure of microarray data. *BMC Bioinformatics.* 2005; 6:1–11. [PubMed: 15631638]
- Qu L, Nettleton D, Dekkers J. Improved estimation of the noncentrality parameter distribution from a large number of t-statistics, with applications to false discovery rate estimation in microarray data analysis. *Biometrics.* 2012; 68:1178–1187. [PubMed: 22551000]
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2013. <http://www.Rproject.org>
- Smyth, GK. Limma: linear models for microarray data. In: Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W., editors. *Bioinformatics and computational biology solutions Using R and bioconductor*. Springer; New York: 2005. p. 397–420.
- Storey J, Taylor J, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B.* 2004; 66:187–205.
- Storey JD. A direct approach to false discovery rates. *J. R. Stat. Soc. B.* 2002; 64:479–498.
- Tibshirani R. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics.* 2006; 7:106. [PubMed: 16512900]
- van Iterson M, van de Wiel M, Boer J, de Menezes R. General power and sample size calculations for high-dimensional genomic data. *Stat. Appl. Genet. Mol. Biol.* 2013; 4:449–467. [PubMed: 23934609]
- Zehetmayer S, Bauer P, Posch M. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics.* 2005; 21:3771–3777. [PubMed: 16091414]
- Zehetmayer S, Posch M. Post hoc power estimation in large-scale multiple testing problems. *Bioinformatics.* 2010; 26:1050–1056. [PubMed: 20189938]

Figure 1.

Theoretical operation characteristic curves: (Contingent) Power as a function of n where the actual values of $\mu=0$ and σ are used for the power calculations based on (1) and (2). Results are given for $\Delta=0$ (left) and $\Delta=0.1$ (right) with $m=1000$, $\pi_0=0.9$ and $\alpha=0.05$. Each plot shows results for $\sigma \in \{0.8, 1, 1.2, 1.5\}$, indicated by the different gray colors (black line for $\sigma=0.8$, ..., light gray line for $\sigma=1.5$).

Figure 2.

Distribution of total sample size and proportion of rejected alternative hypotheses: The boxplots of the left column show the distribution of the overall sample size for different values of π_0 and n_1 . The right column shows the distribution of the proportion of rejected alternative hypotheses with effect sizes exceeding δ . For each scenario and value of n_1 two boxplots are shown, the left for $\delta=0$ (overall power), the right for $\delta=0.1$ (contingent power). The simulations are performed for independent test statistics with $m=10,000$, $\sigma=1$, and $\alpha=0.05$.

Table 1

Adaptive design with sample size reassessment for $m=1000$ one-sample hypothesis tests and independent test statistics.

σ	π_0	n_1	$=0$			$=0.1$		
			n (sd)	Power	FDR	n (sd)	Power	FDR
0.8	0.8	10	108 (14.4)	0.75	0.04	66 (6.6)	0.75	0.04
$n_0^* = 171$		30	151 (16.9)	0.79	0.04	84 (6.7)	0.79	0.04
$n_{\Delta}^* = 90$		50	162 (17.4)	0.79	0.04	88 (6.8)	0.80	0.04
	0.9	10	129 (20.9)	0.75	0.05	77 (9.5)	0.75	0.05
$n_0^* = 202$		30	185 (29.1)	0.79	0.04	101 (11.2)	0.79	0.04
$n_{\Delta}^* = 106$		50	194 (30.2)	0.80	0.04	104 (11.3)	0.80	0.04
	0.99	10	267 (66.1)	0.78	0.05	136 (25.4)	0.77	0.05
$n_0^* = 307$		30	331 (85.5)	0.80	0.05	197 (72.9)	0.82	0.05
$n_{\Delta}^* = 159$		50	329 (84.6)	0.80	0.05	195 (70.1)	0.82	0.05
1	0.8	10	78 (10.5)	0.76	0.04	51 (5.3)	0.76	0.04
$n_0^* = 111$		30	102 (11.3)	0.79	0.04	63 (5.2)	0.79	0.04
$n_{\Delta}^* = 66$		50	107 (10.8)	0.80	0.04	65 (4.9)	0.80	0.04
	0.9	10	97 (16.0)	0.76	0.05	62 (7.8)	0.77	0.05
$n_0^* = 131$		30	124 (19.8)	0.80	0.04	75 (8.8)	0.80	0.04
$n_{\Delta}^* = 77$		50	129 (20.1)	0.80	0.04	77 (8.8)	0.80	0.04
	0.99	10	230 (68.7)	0.81	0.05	121 (26.5)	0.80	0.05
$n_0^* = 199$		30	263 (100.8)	0.82	0.05	150 (60.0)	0.83	0.05
$n_{\Delta}^* = 116$		50	257 (100.2)	0.81	0.05	147 (57.8)	0.82	0.05
1.2	0.8	10	59 (7.2)	0.77	0.04	41 (4.0)	0.77	0.04
$n_0^* = 78$		30	74 (8.1)	0.79	0.04	49 (4.2)	0.79	0.04
$n_{\Delta}^* = 51$		50	76 (7.6)	0.80	0.04	51 (2.6)	0.80	0.04
	0.9	10	73 (12.5)	0.77	0.04	50 (6.6)	0.77	0.05
$n_0^* = 92$		30	88 (13.4)	0.80	0.05	58 (6.7)	0.80	0.05
$n_{\Delta}^* = 59$		50	92 (14.0)	0.80	0.05	60 (6.7)	0.79	0.04

σ	π_0	n_1	$=0$			$=0.1$		
			n (sd)	Power	FDR	n (sd)	Power	FDR
$n_0^* = 140$	0.99	10	195 (68.1)	0.83	0.05	107 (27.4)	0.82	0.05
		30	200 (95.6)	0.82	0.05	116 (47.8)	0.83	0.05
		50	194 (92.1)	0.82	0.05	114 (46.4)	0.82	0.05
$n_{\Delta}^* = 90$	1.5	10	41 (4.9)	0.77	0.04	31 (3.0)	0.77	0.04
		30	50 (5.1)	0.80	0.04	36 (3.0)	0.80	0.04
		50	52 (3.7)	0.79	0.04	50 (0.0)	0.84	0.04
$n_0^* = 61$	0.9	10	52 (9.2)	0.78	0.05	38 (5.4)	0.78	0.05
		30	59 (8.7)	0.80	0.05	42 (4.9)	0.80	0.05
		50	61 (8.2)	0.80	0.05	50 (1.1)	0.82	0.04
$n_{\Delta}^* = 42$	0.99	10	151 (62.9)	0.84	0.05	89 (27.1)	0.84	0.05
		30	136 (76.5)	0.82	0.05	85 (36.9)	0.83	0.05
		50	131 (71.8)	0.82	0.05	84 (34.1)	0.82	0.05

The table shows the simulated average sample sizes, their standard deviations, the actual (contingent) power, the actual FDR and the oracle sample size n_{Δ}^* , derived from (1) given the true π_0 and true distribution of effect sizes under the alternative. Power is defined as the proportion of rejected alternative hypotheses with effect size $>$ under all such alternative hypotheses. For $=0$ this amounts to the overall power.

Table 2

Adaptive design with sample size reassessment for $m=10,000$ one-sample hypothesis tests and independent test statistics.

σ	π_0	n_1	$=0$			$=0.1$		
			n (sd)	Power	FDR	n (sd)	Power	FDR
0.8	0.8	10	105 (4.4)	0.74	0.04	65 (2.1)	0.74	0.04
$n_0^* = 171$		30	149 (5.0)	0.79	0.04	84 (2.0)	0.79	0.04
$n_{\Delta}^* = 90$		50	159 (5.3)	0.79	0.04	87 (2.1)	0.80	0.04
	0.9	10	126 (6.1)	0.74	0.04	76 (2.8)	0.74	0.05
$n_0^* = 202$		30	180 (8.8)	0.79	0.04	99 (3.4)	0.79	0.04
$n_{\Delta}^* = 106$		50	189 (8.7)	0.79	0.05	103 (3.4)	0.80	0.05
	0.99	10	258 (23.6)	0.78	0.05	132 (8.2)	0.77	0.05
$n_0^* = 307$		30	334 (49.0)	0.81	0.05	169 (19.8)	0.81	0.05
$n_{\Delta}^* = 159$		50	320 (48.5)	0.80	0.05	165 (19.1)	0.80	0.05
1	0.8	10	76 (3.1)	0.76	0.04	51 (1.7)	0.76	0.04
$n_0^* = 111$		30	101 (3.3)	0.79	0.04	62 (1.5)	0.79	0.04
$n_{\Delta}^* = 66$		50	106 (3.4)	0.79	0.04	65 (1.7)	0.80	0.04
	0.9	10	90 (4.6)	0.76	0.05	59 (2.3)	0.76	0.05
$n_0^* = 131$		30	118 (5.7)	0.79	0.05	72 (2.7)	0.79	0.05
$n_{\Delta}^* = 77$		50	123 (5.9)	0.79	0.05	74 (2.6)	0.79	0.04
	0.99	10	215 (24.5)	0.81	0.05	117 (9.2)	0.80	0.05
$n_0^* = 199$		30	220 (39.0)	0.81	0.05	125 (15.6)	0.81	0.05
$n_{\Delta}^* = 116$		50	209 (34.8)	0.80	0.05	121 (14.4)	0.80	0.05
1.2	0.8	10	58 (2.5)	0.77	0.04	40 (1.3)	0.77	0.04
$n_0^* = 78$		30	72 (2.5)	0.79	0.04	48 (1.3)	0.79	0.04
$n_{\Delta}^* = 51$		50	76 (2.4)	0.80	0.04	50 (0.3)	0.80	0.04
	0.9	10	70 (3.7)	0.77	0.04	49 (2.0)	0.77	0.05
$n_0^* = 92$		30	86 (4.2)	0.79	0.04	57 (2.2)	0.79	0.05
$n_{\Delta}^* = 59$		50	89 (4.2)	0.80	0.05	58 (2.1)	0.80	0.04

σ	π_0	n_1	$=0$			$=0.1$		
			n (sd)	Power	FDR	n (sd)	Power	FDR
$n_0^* = 140$	0.99	10	176 (23.7)	0.82	0.05	101 (9.7)	0.82	0.05
		30	153 (26.4)	0.81	0.05	95 (12.1)	0.81	0.05
		50	147 (23.0)	0.80	0.05	93 (10.9)	0.80	0.05
$n_{\Delta}^* = 90$	0.8	10	40 (1.5)	0.77	0.04	30 (1.1)	0.77	0.04
		30	49 (1.5)	0.79	0.04	35 (0.8)	0.80	0.04
		50	50 (0.8)	0.80	0.04	50 (0.0)	0.84	0.04
$n_0^* = 61$	0.9	10	50 (2.7)	0.78	0.04	37 (1.7)	0.78	0.05
		30	58 (2.7)	0.80	0.05	42 (1.5)	0.80	0.04
		50	59 (2.8)	0.80	0.04	50 (0.0)	0.82	0.04
$n_{\Delta}^* = 42$	0.99	10	129 (20.3)	0.83	0.05	80 (9.4)	0.83	0.05
		30	100 (16.6)	0.81	0.05	68 (8.8)	0.81	0.05
		50	96 (14.5)	0.80	0.05	67 (7.9)	0.80	0.05

See Table 1 for the legend.

Table 3

Adaptive design with sample size reassessment for unimodal effect size distribution for two-sample t-tests for the mixfdr and the SSPA method.

m	π_0	n_1	$=0$				$=0.1$			
			mixfdr		SSPA		mixfdr		SSPA	
			n (sd)	Power	n (sd)	Power	n (sd)	Power	n (sd)	Power
1000	0.8	10	140 (16.5)	0.75	38 (19.9)	0.50	93 (8.7)	0.75	37 (20.8)	0.54
$n_0^* = 215$		30	191 (20.7)	0.79	76 (45.4)	0.63	117 (9.7)	0.79	67 (25.4)	0.68
$n_{\Delta}^* = 127$		50	203 (21.5)	0.79	105 (63.0)	0.68	122 (10.0)	0.79	84 (32.6)	0.72
	0.9	10	176 (29.7)	0.76	64 (66.9)	0.53	114 (14.7)	0.76	61 (75.4)	0.56
$n_0^* = 253$		30	232 (36.9)	0.79	132 (103.5)	0.67	139 (16.4)	0.79	105 (69.6)	0.71
$n_{\Delta}^* = 150$		50	244 (37.9)	0.79	172 (121.7)	0.71	146 (17.6)	0.80	135 (78.7)	0.75
	0.99	10	343 (77.8)	0.79	217 (151.8)	0.66	233 (76.7)	0.79	192 (148.2)	0.69
$n_0^* = 359$		30	355 (71.7)	0.79	363 (82.4)	0.78	265 (88.4)	0.82	343 (89.1)	0.85
$n_{\Delta}^* = 226$		50	359 (67.3)	0.79	384 (54.2)	0.80	265 (86.2)	0.82	380 (56.5)	0.86
10,000	0.8	10	140 (5.3)	0.75	32 (3.8)	0.48	93 (2.7)	0.75	32 (3.8)	0.52
$n_0^* = 215$		30	188 (6.5)	0.79	60 (10.1)	0.62	116 (3.1)	0.79	60 (9.9)	0.67
$n_{\Delta}^* = 127$		50	201 (6.6)	0.79	77 (15.5)	0.66	121 (3.2)	0.79	76 (12.6)	0.71
	0.9	10	172 (8.1)	0.76	35 (7.0)	0.46	111 (4.4)	0.76	35 (6.4)	0.50
$n_0^* = 253$		30	225 (11.0)	0.79	78 (23.0)	0.63	137 (5.2)	0.79	75 (19.6)	0.68
$n_{\Delta}^* = 150$		50	237 (11.7)	0.79	104 (37.2)	0.67	142 (5.2)	0.79	98 (30.4)	0.72
	0.99	10	354 (47.1)	0.79	70 (82.4)	0.45	217 (36.6)	0.80	68 (83.2)	0.49
$n_0^* = 359$		30	373 (35.7)	0.79	280 (121.7)	0.74	226 (29.9)	0.80	259 (117.3)	0.79
$n_{\Delta}^* = 226$		50	373 (35.0)	0.80	345 (90.5)	0.78	224 (25.7)	0.80	351 (82.4)	0.85

See the legend of Table 1. All sample sizes are per-group sample sizes. The oracle sample size n_{Δ}^* was determined by simulating from the true distribution of effect sizes.

Table 4

Microarray-based simulation with unimodal ($\sigma=1.5$) and bimodal effect size distribution for two-sample t-tests.

π_0	n_1	$=0$				$=0.1$			
		mixfdr		SSPA		mixfdr		SSPA	
		n (sd)	Power	n (sd)	Power	n (sd)	Power	n (sd)	Power
Unimodal effect size distribution									
0.8	10	68 (7.4)	0.76	40 (23.0)	0.65	50 (4.5)	0.76	39 (23.1)	0.68
$n_0^* = 98$	30	87 (7.0)	0.79	77 (37.2)	0.75	62 (3.9)	0.79	75 (37.4)	0.78
$n_\Delta^* = 68$	50	92 (6.2)	0.79	91 (35.4)	0.78	65 (3.1)	0.80	89 (35.7)	0.82
0.9	10	81 (12.8)	0.76	72 (36.4)	0.72	59 (7.4)	0.76	69 (34.2)	0.75
$n_0^* = 116$	30	102 (10.8)	0.79	111 (31.6)	0.78	73 (7.1)	0.79	107 (34.8)	0.82
$n_\Delta^* = 79$	50	108 (8.8)	0.79	115 (26.6)	0.79	76 (5.8)	0.80	112 (29.2)	0.82
0.99	10	120 (13.4)	0.76	104 (27.9)	0.73	105 (21.5)	0.78	104 (27.9)	0.77
$n_0^* > 127$	30	126 (4.7)	0.76	127 (1.0)	0.76	109 (14.0)	0.79	127 (1.0)	0.81
$n_\Delta^* = 121$	50	127 (1.9)	0.76	127 (0.2)	0.76	113 (12.0)	0.79	127 (0.2)	0.81
$=0$					$=0.5$				
Bimodal effect size distribution									
0.8	10	23 (3.4)	0.70	36 (16.0)	0.81	11 (1.4)	0.32	29 (14.9)	0.78
$n_0^* = 30$	30	34 (8.2)	0.83	62 (31.9)	0.91	30 (0.0)	0.83	31 (10.8)	0.84
$n_\Delta^* = 28$	50	51 (6.9)	0.92	82 (34.4)	0.95	50 (0.0)	0.94	51 (9.0)	0.94
0.9	10	29 (6.2)	0.70	68 (33.3)	0.90	11 (1.4)	0.26	49 (32.3)	0.83
$n_0^* = 35$	30	42 (19.3)	0.82	103 (35.7)	0.95	30 (0.0)	0.78	50 (34.0)	0.86
$n_\Delta^* = 32$	50	56 (17.4)	0.90	109 (31.1)	0.96	50 (0.0)	0.92	63 (31.5)	0.94
0.99	10	86 (38.2)	0.87	105 (28.2)	0.93	51 (47.4)	0.47	81 (35.1)	0.88
$n_0^* = 51$	30	84 (41.6)	0.84	127 (0.3)	0.96	48 (39.9)	0.68	94 (28.5)	0.94
$n_\Delta^* = 46$	50	81 (35.2)	0.87	127 (0.0)	0.96	60 (28.5)	0.85	90 (34.5)	0.92

The table shows the simulated average sample sizes and their standard deviations and the actual (contingent) power for the mixfdr and the SSPA method. The oracle sample size n_Δ^* was determined by simulating from the true distribution of effect sizes.

Table 5

Adaptive design with sample size reassessment for bimodal effect size distribution for two-sample t-tests (microarray simulation).

n_1	$=0$				$=0.5$			
	<u>mixfdr</u>		<u>SSPA</u>		<u>mixfdr</u>		<u>SSPA</u>	
	<i>n</i>	Power	<i>n</i>	Power	<i>n</i>	Power	<i>n</i>	Power
10	26 (7)	0.68	67 (32.9)	0.89	18 (26.2)	0.31	47 (33.4)	0.83
30	45 (18.4)	0.83	105 (32.3)	0.95	30 (0.0)	0.78	49 (29.6)	0.86
50	56 (22.3)	0.90	107 (33.9)	0.96	50 (0.0)	0.92	64 (32.0)	0.93

1/4 of the alternatives have an effect size distribution with negative mean, 3/4 with positive mean. The table shows the simulated average sample sizes and the actual (contingent) power for the mixfdr and the SSPA method ($\pi_0=0.9$, $m=1000$).

Table 6

Adaptive design with sample size reassessment for two-sample t-tests.

	n_1	$=0$				$=0.5$			
		mixfdr		SSPA		mixfdr		SSPA	
		n	Power	n	Power	n	Power	n	Power
Bimodal effect size distribution									
$n_0^* = 35$	10	127 (0.5)	0.98	66 (32.8)	0.89	36 (1.8)	0.84	46 (32.8)	0.83
$n_\Delta^* = 32$	30	127 (0.0)	0.98	105 (34.1)	0.95	41 (1.1)	0.88	51 (33.6)	0.86
	50	127 (0.0)	0.98	110 (30.3)	0.96	50 (0.0)	0.92	64 (29.7)	0.93
Unimodal effect size distribution									
$n_0^* = 116$	10	59 (31.8)	0.68	75 (35.6)	0.73	16 (25.4)	0.45	50 (37.2)	0.85
$n_\Delta^* = 26$	30	73 (26.0)	0.74	113 (30.8)	0.79	30 (6.5)	0.81	56 (35.8)	0.88
	50	62 (18.1)	0.72	115 (26.7)	0.79	50 (0.0)	0.91	63 (29.5)	0.92

Simulations under a bimodal (upper table) and unimodal (lower table) effect size distribution applying the mixfdr method fitting a misspecified model, i.e. a unimodal (upper table) or bimodal (lower table) distribution for the effect sizes under the alternative. For comparison also the results for the SSPA method are given ($\pi_0=0.9$, $m=1000$).

Table 7

Adaptive design with sample size reassessment for unimodal effect size distribution for two-sample t-tests.

n_1	$\delta=0$		$\delta=0.1$		$\delta=0.3$		$\delta=0.5$		$\delta=1$	
	n	Power	n	Power	n	Power	n	Power	n	Power
mixfdr										
10	177 (29.0)	0.76	114 (14.7)	0.75	60 (5.6)	0.76	38 (2.9)	0.77	19 (1.2)	0.77
30	231 (36.2)	0.79	139 (16.6)	0.79	68 (5.5)	0.8	43 (2.9)	0.8	30 (0)	0.94
50	243 (37.5)	0.8	145 (16.8)	0.8	70 (5.9)	0.80	50 (0.1)	0.85	50 (0)	1
SSPA										
10	69 (70.0)	0.53	61 (68.4)	0.56	48 (61.2)	0.62	36 (33.6)	0.69	18 (3.5)	0.73
30	132 (102.8)	0.67	103 (64.4)	0.71	70 (55.0)	0.77	49 (38.1)	0.8	30 (0)	0.94
50	177 (125.5)	0.71	125 (81.0)	0.75	75 (45.3)	0.80	52 (18.0)	0.85	50 (0)	1

The table shows the simulated average sample sizes and the actual (contingent) power for the mixfdr and the SSPA method for $\pi_0=0.9$, $m=1000$, and different values of δ .

Table 8

Application to a microarray gene expression study applying the mixfdr method for two-sample t-tests.

n_{1A}/n_{1B}	$\pi=0$			$\pi=0.1$		
	Estimated interim power	Estimated power	n_A/n_B	Estimated interim power	Estimated power	n_A/n_B
Control/Interstitial lung disease						
10/20	0.29 (0.96, 0.64)	0.76 (0.78, 0.30)	108/216	0.31	0.80 (0.77, 0.30)	86/172
30/60	0.58 (0.89, 0.40)	0.76 (0.79, 0.29)	108/216	0.64	0.80 (0.79, 0.29)	83/166
50/100	0.67 (0.85, 0.30)	0.76 (0.80, 0.29)	108/216	0.74	0.80 (0.68, 0.29)	85/170
Cpd/Interstitial lung disease						
10/10	0.28 (1.10, 0.52)	0.72 (0.70, 0.41)	141/141	0.30	0.70 (0.66, 0.42)	87/87
30/30	0.42 (0.71, 0.43)	0.76 (0.65, 0.31)	220/220	0.47	0.81 (0.67, 0.41)	157/157
50/50	0.52 (0.68, 0.43)	0.76 (0.66, 0.30)	220/220	0.58	0.82 (0.67, 0.37)	167/167
Control/Cpd						
10/20	0.25 (1.08, 0.89)	0.43 (0.35, 0.56)	108/216	0.27	0.52 (0.35, 0.57)	98/196
30/60	0.44 (0.64, 0.44)	0.43 (0.35, 0.57)	108/216	0.50	0.54 (0.35, 0.57)	108/216
50/100	0.51 (0.59, 0.48)	0.40 (0.34, 0.59)	108/216	0.58	0.52 (0.34, 0.58)	108/216

The estimated (interim) power, the estimated parameters of the effect size distribution in parentheses (σ and π_0) and the resulting n for both groups are reported.