Two-stage designs for experiments with a large number of hypotheses

Sonja Zehetmayer, Peter Bauer and Martin Posch*

Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

Received on April 26, 2005; revised on July 1, 2005; accepted on July 28, 2005

Advance Access publication August 9, 2005

ABSTRACT

Motivation: When a large number of hypotheses are investigated the false discovery rate (FDR) is commonly applied in gene expression analysis or gene association studies. Conventional single-stage designs may lack power due to low sample sizes for the individual hypotheses. We propose two-stage designs where the first stage is used to screen the ‘promising’ hypotheses which are further investigated at the second stage with an increased sample size. A multiple test procedure based on sequential individual \( P \)-values is proposed to control the FDR for the case of independent normal distributions with known variance.

Results: The power of optimal two-stage designs is impressively larger than the power of the corresponding single-stage design with equal costs. Extensions to the case of unknown variances and correlated test statistics are investigated by simulations. Moreover, it is shown that the simple multiple test procedure using first stage data for screening purposes and deriving the test decisions only from second stage data is a very powerful option.

Availability: An R-program is available at http://www.meduniwien.ac.at/medstat/research/fdr/application.R

Contact: Martin.Posch@meduniwien.ac.at

Supplementary information: Supplementary data for this paper is available at Bioinformatics online.

1 INTRODUCTION

In gene expression and gene association studies, typically a large number of hypotheses tests are performed, but only a small percentage is expected to show an effect. As there are many tests but only a small percent-

age is expected to show an effect, there is a trade off between the number of selected hypotheses and the sample size for each hypothesis is preplanned. However, the second stage sample size is divided among the selected hypotheses. In contrast to Satagopan et al. (2002) we do not select a prefixed number of hypotheses but all hypotheses whose univariate first stage \( P \)-values are below a certain boundary, i.e. to reject at least one true null hypothesis. However, in situations with a large number of hypotheses, the control of the FWE leads to conservative procedures with a low power to identify the few existing effects. A less conservative approach for the multiple testing problem is to control the false discovery rate (FDR), see e.g. Benjamini and Hochberg (1995). The FDR controls the expected proportion of type I errors among the rejected hypotheses. A number of recent articles deal with multiple testing in classical single-stage designs (Dudoit et al., 2003; Reiner et al., 2003) that control the FWE or the FDR. Given a fixed overall sample size, Futschik and Posch (2005) showed that efficiency (defined as the expected number of detected effects) can be gained by randomly selecting a smaller number of hypotheses such that more observations for each hypothesis are available.

In extension of the single-stage design two types of two-stage designs have been proposed. In the first approach, stage wise sample sizes for each hypothesis are preplanned. However, the second stage data is collected only for a limited number of hypotheses for which the first stage data showed promising effects. Thus, the total number of observations (across stages and hypotheses) is random. Following this idea, Miller et al. (2001) advocated a two-stage design for gene expression experiments. They propose to use the first stage data only for the selection of hypotheses. To control the FWE in the second stage for the selected hypotheses a Bonferroni test is performed using only the second stage data. Satagopan and Elston (2003) improved this procedure by using group sequential methods to incorporate the first stage data in the final Bonferroni test. Both approaches are very conservative as they rely on Bonferroni adjusted critical values. Very recently an FDR-controlling two-stage design using the concept of FDR for selecting at the first stage and for confirmation at the second stage has been proposed by Benjamini and Yekutieli (2005).

In the second type of two-stage designs it is assumed that the overall number of observations (or more generally total costs) is fixed. A certain fraction of these observations is spent in the first stage. The remaining observations are then distributed among the hypotheses selected for the second stage. In this approach, the second stage sample size for each hypothesis is random. Satagopan et al. (2002) applied this idea in the context of gene association studies. In their procedure only a small prefixed number of hypotheses, which are determined by the smallest univariate \( P \)-values, is rejected in the final test. This procedure neither controls the FWE nor the FDR.

Extending these approaches, we propose two-stage designs controlling the FDR, where, based on first stage data, hypotheses are selected for the second stage. In contrast to Satagopan et al. (2002) we do not select a prefixed number of hypotheses but all hypotheses whose univariate first stage \( P \)-values lie below a certain boundary. We also assume that there is a fixed total sample size (or, more generally, fixed total costs), which is split between the two stages, and the second stage sample size is divided among the selected hypotheses. Thus, there is a trade off between the number of selected hypotheses and the sample size for each hypothesis. The final test decision is based on data from both stages. In Section 2, the definition of the FDR and an estimator based on \( P \)-values proposed in the literature are outlined. In Section 3, appropriate \( P \)-values are defined for the two-stage test procedure. Here it is assumed that the observations are independently normally distributed between hypotheses with common, known variance. In Section 4 optimal two-stage procedures controlling the FDR and maximizing the expected number of correct
rejections are derived. It is assumed that under the alternative the test statistics follow independent and identical normal distributions. It turns out that such two-stage designs provide a substantial advantage with regard to the probability of correct rejections as compared with the single-stage test procedure. In Section 5 several extensions are investigated. First, optimal designs are considered, when sampling costs differ between stages. Simulations are reported when the procedure is applied in the situation of unknown variances that may differ between the hypotheses. A further extension investigated is that the test statistics under the alternative, instead of having a single mean, have means arising from a gamma distribution. Finally, the statistical properties are simulated for correlated test statistics. In Section 6, a simple procedure is considered, where the first stage is only used for selecting the promising hypotheses to be investigated at the second stage. The FDR is estimated for the set of second stage hypotheses based on the second stage sample only. Section 7 gives some concluding remarks.

2 ESTIMATING THE FDR

Consider a simultaneous test of \( m \) one null hypotheses \( H_{0i}, i = 1, \ldots, m \). The FDR is defined as the expected fraction of erroneously rejected null hypotheses in all rejected null hypotheses. More formally, let \( R \) denote the number of rejected null hypotheses and \( V \) the number of erroneously rejected null hypotheses. Then the FDR is given by

\[
\text{FDR} = \frac{E\left(\frac{V}{R}\right)}{P(R > 0)}. \tag{1}
\]

Assume that the \( m \) identical null hypotheses are tested by their \( P \)-values \( p_i, i = 1, \ldots, m \), at some level \( \gamma \). Storey (2002) proposes an estimator for the resulting FDR. First, the fraction \( \pi_0 \) of true null hypotheses in all \( m \) hypotheses is estimated by

\[
\hat{\pi}_0 = \frac{\text{#}(p_i > \lambda)}{((1 - \lambda)m_1)}, \tag{2}
\]

where \( \lambda \) is a constant chosen a priori and \( \text{#}(p_i > \lambda) \) denotes the number of \( P \)-values exceeding \( \lambda \). Note that increasing \( \lambda \) reduces the bias of \( \hat{\pi}_0 \) at the cost of a higher variance. Now the estimator of the FDR is given by

\[
\hat{\text{FDR}}_\gamma(\gamma) = \frac{\hat{\pi}_0 \gamma m_1}{\max(2\hat{\pi}_0, 1)}. \tag{3}
\]

Storey et al. (2004) showed that the expected value of this estimate is an upper bound for the true false discovery rate if the \( P \)-values corresponding to the true null hypotheses are independent and uniformly distributed. To obtain a test with a specified FDR, the largest \( \gamma \) is determined such that \( \hat{\text{FDR}}_\gamma(\gamma) \leq \alpha \), using the \( P \)-values observed in the sample.

3 TWO-STAGE DESIGNS

3.1 The test problem

We consider \( m \), one-sided hypotheses for the mean of independent, normally distributed observations with known variance \( \sigma^2 \), assuming also independence across hypotheses. We test the hypotheses

\[
H_{0i} : \mu = 0 \quad \text{against} \quad H_{1i} : \mu > 0, \quad i = 1, \ldots, m_1.
\]

Later on we investigate situations where the assumptions of known variance and independence among the observations across hypotheses are relaxed. Also the two-sided case will be investigated.

3.2 The test procedure

Assume there is an overall number \( N \) of available observations. In the first stage a fraction \( r \) of the \( N \) observations is distributed equally (up to round off errors) among the \( m \) hypotheses, the first stage sample size per hypothesis being \( n_1 = rN/m_1 \). Let \( z_i^{(1)} \) denote the standardized first stage mean of the observations for hypothesis \( i \). Then the first stage \( P \)-values are given by \( p_i^{(1)} = 1 - \Phi(z_i^{(1)}) \). After the second stage of the observations, the first stage hypotheses are selected for the second stage. For all others, \( H_{0i} \) is accepted. We denote the random number of selected hypotheses to be carried over to the second stage by \( m_2 \) and the set of selected hypotheses by \( i_1 \cdots i_{m_2} \). At the second stage, the remaining \( (1 - r)N \) observations are equally distributed among the selected \( m_2 \) hypotheses. Thus, the second stage sample size for each selected hypothesis is

\[
\lambda_2 = (1 - r)N/m_2.
\]

Let \( \gamma_1 \) denote the standardized mean of the total sample from both stages for hypothesis \( i \) and let \( p_i = 1 - \Phi(z_i) \) denote the \( P \)-value from the pooled sample after the second stage. Then, for all \( i \) in \( i_1 \cdots i_{m_2}, H_{0i} \) is rejected in the final test if \( p_i < \gamma_2 \) for some constant \( \gamma_2 \). In the following we show how to choose \( \gamma_2 \) to control the FDR at some specified value. To this end, we reformulate the test procedure in terms of an overall \( P \)-value for each of the sequential two-stage tests. The crucial point is that the independent increment structure of group sequential designs is preserved in our situation, where the second stage sample size \( n_2 \) is a random variable.

3.3 A \( P \)-value for the two-stage design

Let us first assume that \( n_2 \) is deterministic. Then the local level of the two-stage test is given by

\[
\gamma = P_{H_{0i}}(p_i^{(1)} \leq \gamma_1, p_i \leq \gamma_2) = \int_{-\infty}^{\gamma_2} \left[ 1 - \Phi\left( \frac{c_1 - \gamma_1 - \sqrt{n_1/m_1} z}{\sqrt{n_1/m_1}} \right) \right] \phi(z) \, dz, \tag{4}
\]

where \( n = n_1 + n_2 \) and \( c_1 = q_{1-\xi} \) denotes the \((1 - \xi)\)-quantile and \( \phi(z) \) the density of the standard normal distribution, respectively.

In the two-stage test introduced above, the second stage sample size is a random variable. However, the conditional distribution of \( n_2 \), given that the \( i \)-th hypothesis is selected (i.e. \( p_i^{(1)} \leq \gamma_1 \)), is independent of \( p_i^{(1)} \). This follows from the assumption of independence of the observations across hypotheses. Hence, (4) gives also the level of the two-stage test if \( n_2 \) is a random variable, as in the above two-stage procedure.

Now, an overall \( P \)-value for the group sequential two-stage test based on a monotonic ordering of the sample space as proposed in Tsiatis et al. (1984) is given by

\[
P_{ad} = \sum_{i_1}^{n_1} \left[ 1 - \Phi\left( \frac{c_1 - \gamma_1 - \sqrt{n_1/m_1} z}{\sqrt{n_1/m_1}} \right) \right] \phi(z) \, dz \quad \text{if} \quad p_i^{(1)} > \gamma_1 \tag{5}
\]

Note that the integral in (5) is the same as in (4) with \( c_1 \) replaced by the observed \( Z \)-statistics in the total sample. With (4), every
critical region \( p_1^{(1)} \leq \gamma_1, p_2 \leq \gamma_2 \) for the overall \( P \)-value in the total sample corresponds to a critical region \( p_2 \leq \gamma \) for the sequential \( P \)-value and vice versa. Additionally, the \( P \)-value \( p_\mu \) is uniformly distributed under \( H_\mu \). To see this, let \( \gamma, 0 \leq \gamma \leq 1 \), be fixed and \( \gamma_2 \) be the solution of (4). Then,

\[
P_{H_\mu}(p_\mu \leq \gamma) = P_{H_\mu}(p_1^{(1)} \leq \gamma_1, p_2 \leq \gamma_2) = \gamma.
\]

### 3.4 Control of the false discovery rate

For given \( \gamma_1 \) and \( \gamma_2 \) the FDR can be estimated with the estimator (3), where the \( p_i \) are replaced by the sequential \( P \)-values, \( p_\mu \). In the appendix in the supplementary data we show that the results of Storey (2002) on the consistency and conservativeness of the estimator of the FDR apply.

To control the FDR at a specified level \( \alpha \), we rewrite the estimate (3) and set

\[
\alpha = \frac{\hat{\pi}_0 m_1 \gamma_2}{\max(\{p_1^{(1)} \leq \gamma_1, p_1 \leq \gamma_2\}),}
\]

where \( \gamma \) as function of \( \gamma_2 \) is given by (4) and \( \hat{\pi}_0 \) is estimated by (2) with \( p_i \) replaced by \( p_\mu \). Now, (6) is solved for \( \gamma_2 \). If all hypotheses \( i \) that have been selected for the second stage and for which \( p_i \leq \gamma_2 \) are rejected, the FDR is controlled (at least asymptotically) at the specified level. An R-program (R Development Core Team, 2005) to apply the procedure to a dataset is available at http://www.meduniwien.ac.at/medstat/research/fdr/application.R

### 4 OPTIMAL TWO-STAGE PROCEDURES

#### 4.1 Asymptotically optimal designs

Given an FDR \( \alpha \), an initial number of hypotheses \( m_1 \) and an overall number of observations \( N \), the two-stage procedure involves two design parameters: the futility bound \( \gamma_1 \) and the fraction of observations to be spent in the first stage \( r \), which determines the first stage sample size \( m_1 = rN/m_1 \). Thus, for a specified alternative we can optimize these parameters with respect to the power, defined as the probability to reject a null hypothesis given the alternative holds.

Assume that for all alternative hypotheses the same alternative \( \mu = \Delta > 0 \) holds. Asymptotically, for a large number of hypotheses and up to round off errors, \( m_2 = m_1[\pi_0 \gamma_1 + (1 - \pi_0) P_{H_\mu = \Delta}(p_1^{(1)} \leq \gamma_1)] \) and \( m_2 = (1 - r)N/m_2 \). Thus, asymptotically the rejection boundary \( \gamma \) for the \( P \)-values in the final analysis of the two-stage procedure is given by the solution of

\[
\alpha = \frac{\pi_0 \gamma}{\pi_0 \gamma + (1 - \pi_0)(1 - \beta(\gamma))},
\]

where

\[
1 - \beta(\gamma) = P_{H_\mu = \Delta}(p_\mu \leq \gamma) = \int_{-\infty}^{\infty} \left[ 1 - \Phi_{\frac{\Delta}{\sqrt{m_1/N}}} \left( \frac{1 - \gamma_1 - \sqrt{m_1/m_2}}{\sqrt{m_2/N}} \right) \right] \psi_{\frac{\Delta}{\sqrt{m_1/N}}} \Phi_{\frac{\Delta}{\sqrt{m_2/N}}} \left( \frac{z}{\sqrt{m_2/N}} \right) dz,
\]

and \( \gamma_2 \) is the solution of (4). Here, \( \Phi_{\mu, \sigma^2} \) and \( \psi_{\mu, \sigma^2} \) are the cumulative distribution and density function of the normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Note that the individual power (8) is equal to the expected proportion of correct rejections in the set of alternative hypotheses. In the following, we refer to this quantity \( \Pi_\alpha = 1 - \beta(\gamma) \) as the ‘power’ of the multiple test procedure.

Now, we optimize the objective function (8) in \( \gamma_1 \) and \( r \), where \( \gamma \) as a function of the targeted FDR \( \alpha \) is implicitly defined by (7). It is easy to see that (8), (4) and (7), and consequently also the optimal \( \gamma_1 \) and \( r \), depend on \( N, m_1, \Delta \) and \( \sigma \) only via \( d = (\Delta/\sigma) \sqrt{N/m_1} \).

Table 1 shows the optimal \( \gamma_1 \) and \( r \) and the resulting power for several scenarios. For comparison also the power of the corresponding single-stage design \( \Pi_0 \) is given. With increasing \( N \), the optimal \( r \) and \( \gamma_1 \) increase slightly. Also for increasing proportions of true null hypotheses, \( r \) increases, whereas the Power \( \Pi_1 \) and the optimal \( \gamma_1 \) decrease. In the considered scenarios (10)–(15) of the \( m_1 \) hypotheses are selected for the second stage and about 2/3 of all available observations are used in the first stage. E.g. for \( \Delta/\sigma = 1, N = 8m_1 \) and \( \pi_0 = 0.99 \), the optimal sample size in the first and second stage is \( m_1 = 5.5 \) and \( m_2 = 19.2 \), and the resulting number of selected hypotheses is \( m_2 = 0.13m_1 \). The optimal sample sizes are similar to those proposed in Satagopan et al.
Table 3. Simulation results

<table>
<thead>
<tr>
<th></th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
<th>FDR</th>
<th>( \Pi_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known variance</td>
<td>0.00045 (0.000032)</td>
<td>0.00071 (0.000055)</td>
<td>0.0488 (0.033)</td>
<td>0.848 (0.051)</td>
</tr>
<tr>
<td>Unknown variance</td>
<td>0.00041 (0.000037)</td>
<td>0.00065 (0.000062)</td>
<td>0.0493 (0.034)</td>
<td>0.774 (0.061)</td>
</tr>
<tr>
<td>Distributed mean</td>
<td>0.00028 (0.000042)</td>
<td>0.00043 (0.000068)</td>
<td>0.0497 (0.042)</td>
<td>0.523 (0.074)</td>
</tr>
<tr>
<td>( \rho ) = 0.20</td>
<td>0.00045 (0.000032)</td>
<td>0.00071 (0.000055)</td>
<td>0.0487 (0.033)</td>
<td>0.848 (0.051)</td>
</tr>
<tr>
<td>( \rho ) = 0.60</td>
<td>0.00045 (0.000034)</td>
<td>0.00072 (0.000057)</td>
<td>0.0491 (0.036)</td>
<td>0.848 (0.051)</td>
</tr>
<tr>
<td>( \rho ) = 0.80</td>
<td>0.00046 (0.000041)</td>
<td>0.00072 (0.000067)</td>
<td>0.0581 (0.050)</td>
<td>0.849 (0.052)</td>
</tr>
<tr>
<td>( \rho ) = 0.90</td>
<td>0.00048 (0.000063)</td>
<td>0.00075 (0.000010)</td>
<td>0.0860 (0.086)</td>
<td>0.851 (0.052)</td>
</tr>
<tr>
<td>( \rho ) = 0.98</td>
<td>0.00052 (0.000018)</td>
<td>0.00083 (0.00028)</td>
<td>0.0968 (0.154)</td>
<td>0.851 (0.061)</td>
</tr>
</tbody>
</table>

Average critical boundaries \( \gamma_1, \gamma_2 \), FDR and power for the known variance case, and the unknown variance case, and gamma distributed means and correlated observations across hypotheses. 100,000 simulation runs have been performed for \( N = 40,000 \), \( m_1 = 5000 \), \( \Delta / \sigma = 1 \), \( \pi_0 = 0.99 \), \( \alpha = 0.05 \), \( \lambda = 0.5 \), \( r = 0.625 \) and \( \gamma_1 = 0.123 \). Standard deviations between the simulation runs are given in parentheses.

Fig. 1. Distribution (minimum, maximum and quartiles) of the actual FDR in the simulated 5000 samples for the scenario \( N = 40,000 \), \( m_1 = 5000 \), \( \Delta / \sigma = 1 \), \( \pi_0 = 0.99 \), \( \alpha = 0.05 \), \( \lambda = 0.5 \), \( r = 0.625 \) and \( \gamma_1 = 0.123 \): (a) known variance case, (b) unknown variance case and (c) gamma distributed means under the alternatives and correlated observations across hypotheses with different values of \( \rho \).

(2002), where the power to select a single (or predefined number of) true alternative(s) is maximized without controlling the type I error. The comparison of the two-stage power and the single-stage power shows very large advantages for the sequential design which slightly decrease with increasing \( N \).

Since the optimal \( r \) and \( \gamma_1 \) depend on \( \Delta / \sigma \) and \( \pi_0 \), which are typically unknown, we investigated the impact of misspecifications in the planning phase. Table 2 shows the power of a two-stage design planned under the assumptions \( N = 8m_1 \), \( \pi_0 = 0.99 \) and \( \Delta / \sigma = 1 \) under different scenarios. Comparison of the power of the optimal two-stage designs in Table 2 shows that the optimum appears to be flat.

4.2 Simulations

To assess finite sample properties we performed a simulation study using the asymptotically optimal design parameters. We modified the optimal \( r \) to get an integer first stage sample size \( n_1 \), rounded to the smaller integer. In the second stage, the value of \( n_2 \) is again rounded to the smaller integer not to exceed the predefined total \( N \). The data was assumed to be independent normally distributed with known variance. The first section in Table 3 shows the results of the simulation study for \( N = 40,000 \), \( m_1 = 5000 \), \( \Delta / \sigma = 1 \), \( \pi_0 = 0.99 \), a targeted FDR of \( \alpha = 0.05 \) and \( \lambda = 0.5 \). In all scenarios the procedure controls the FDR well and is even slightly conservative. The average power for the simulations for integer values of \( n_1 \) and \( n_2 \) is only marginally lower than the analytically optimal power based on non-integer sample sizes. Boxplots for the FDR from the simulations are given in Figure 1a to describe the sampling distribution of the actual FDR. The distribution is skewed with occasional quite large FDR values.

5 EXTENSIONS

5.1 Different sampling costs at the two stages

If the sampling costs vary between the two stages, the total costs are given by \( N = m_1n_1 + Cm_2n_2 \), for some constant \( C \). Then for fixed total costs \( N \), the second stage sample size for each hypothesis is given by \( n_2 = (N - m_1n_1) / (Cm_2) \). Also for this more general setup optimal parameters can be derived as in Section 4. For increasing values of \( C \), less sample size is used for stage two. E.g. for \( N = 8m_1 \), \( \pi_0 = 0.99 \), \( \Delta / \sigma = 1 \) and \( C = 3 \), the optimal power is 0.719 with the parameters \( r = 0.737 \) and \( \gamma_1 = 0.041 \) compared with a power of 0.859 and parameters \( r = 0.674 \) and \( \gamma_1 = 0.138 \) for \( C = 1 \). Thus, while the optimal \( r \), which corresponds to the proportion of total costs to be spent at the first stage, is \( \sim 2/3 \) in both cases, the optimal design selects much fewer hypotheses for the second stage as \( C \) increases. Hence, if, e.g. \( C = 3 \) the total sample size used at the second stage is about one third of the sample size in the scenario with equal sampling costs. The optimal second stage sample size for each selected hypothesis is \( \sim 2/3 \) of the sample size in the scenario with equal sampling costs.

5.2 The \( t \)-test

If the variance is unknown but the same for all hypotheses, the two stage test for the known variance case is still valid because of the large sample size used for the common variance estimate. However, if \( \sigma^2 \) differs between the hypotheses, this approximation is questionable. Since the exact computation of group sequential \( P \)-values is numerically difficult we use an approximation based on the \( P \)-values of the \( t \)-test from the first stage and the pooled sample, denoted by \( P_1^{(1)} \) and \( \tilde{P}_1 \), respectively. The level of the sequential \( t \)-test which rejects...
the distribution gets a large variability and shows a mean above the
increases with increasing correlation. Only for large correlations
rejections of true null hypotheses are rare so that the median FDR
of the FDR also under weak dependence. However, for correla-
with the result of Storey (2004) on the asymptotic control
the sequence of hypotheses.

Using this approximation, \( \gamma_2 \) leading to a specified FDR can be
computed as in Section 3.4.

The performance of the approximations is assessed by simula-
tions where the variances are estimated individually for each null
hypothesis. The optimal parameters \( \pi_1 \) and \( r \) for the corresponding
asymptotic known variance case are used in the simulations, again
using sample sizes rounded to the lower integer. As can be seen from
Table 3, the FDR is well controlled at the specified value and slightly
larger than in the known variance case. As expected, the \( t \)-test has
lower power than the \( Z \)-test. Looking at the boxplots in Figure 1 (b)
the distribution is very similar to the known variance case (a).

5.3 Distributed alternatives

Up to this point we assumed that all alternatives have the same
mean effect. Now we investigate the procedure under the assumption
that, given the alternative holds, the mean effect \( \Delta_1/\sigma \) is distributed
according to a gamma distribution,

\[
\mathbb{P}(x) = \frac{x^{a-1} \exp(-x/b)}{b^a \Gamma(a)}
\]

with the shape parameter \( a = 2 \) and the scale parameter \( b = \Delta_1/\sigma = \frac{1}{2} \),
which leads to a mean of \( ab = 1 \). As shown in Table 3, the power
is smaller than for independently distributed alternatives with the same
mean effect size. The average FDR falls below the targeted 0.05.
Figure 1 shows the distribution of the actual false discovery rates,
which are very similar to the case of fixed alternatives.

5.4 Correlated test statistics

In many testing situations the test statistics are not independent across
hypotheses. To investigate the influence of correlation, we assume an
order among hypotheses and an autoregressive correlation structure.
Hence, the correlation between hypotheses \( i \) and \( j \) is given by \( \rho^{(c_{ij})} \),
for some \( c \in (0, 1) \). The alternatives are randomly distributed among
the sequence of hypotheses.

We assume that the variance is known and the hypotheses have
identical marginal distributions under the alternative. It can be seen
that the power hardly changes with increasing correlation (Table 3).
For low correlation also the FDR is still controlled, which is in line
with the result of Storey et al. (2004) on the asymptotic control
of the FDR also under weak dependence. However, for correlations
\( c > 0.6 \), the procedure becomes anti-conservative. The mean
and median FDR are increasing in \( \rho \). For very large correlation, \( \rho = 0.98 \),
rejections of true null hypotheses are rare so that the median FDR
is zero (Fig. 1). As expected, the variability of the actual FDR increases
with increasing correlation. Only for large correlations the distribution
gets a large variability and shows a mean above the
targeted FDR.

5.5 The two-sided case

Two-sided tests can be constructed by applying two one-sided
multiple tests simultaneously as has been proposed for group sequen-
tial designs (e.g. Jennison and Turnbull, 2000). For the two-stage
approach, the upper one-sided sequential \( P \)-value is computed as in
(5). The lower sequential \( P \)-value is calculated accordingly integ-
rating over the region \( (-\infty, -c_1 \gamma_1) \) and replacing the expression in
the bracket squares by \( \Phi(\tilde{z}) \).

Then one can simply combine the \( 2m_1 \) one-sided hypotheses into
a single set of null hypotheses, using the one-sided \( P \)-values and
proceed as defined in 3.2.

6 THE PILOT DESIGN: IGNORING THE FIRST STAGE DATA FOR THE TEST DECISION

A simple alternative to the sequential two-stage design is to use the
first stage only for the selection of the hypotheses to be continued
to the second stage. Testing is performed only with the observations
from the second stage. We apply the procedure defined in Section 2
to the second stage \( P \)-values \( p_i^2 \) aiming at an FDR of \( \alpha \) and estimate
the FDR by \( \hat{\pi}_0^{(2)} = (2[\{p_i^2 > \lambda\}]/(1 - \lambda)m_2) \) and \( \text{FDR}_{\gamma_2}(\gamma) = \hat{\pi}_0^{(2)} \gamma m_2/ \max(2[\{p_i^2 < \gamma\}], 1) \). Note that \( \hat{\pi}_0^{(2)} \) is an estimate of the
proportion of true null hypotheses among the hypotheses tested in
the second stage. Then, the power of the whole experiment \( \Pi_2 \) is the
product of the first stage power, \( 1 - \beta_1(\gamma_1) = P_{\hat{p}_0 - \Delta_1}(\hat{p}_1^2 \leq \gamma_1) \),
and the second stage power, \( 1 - \beta_2(\gamma_2) = P_{\hat{p}_0 - \Delta_2}(\hat{p}_2^2 \leq \gamma_2) \), since
they are independent. Here, \( \hat{p}_2^{(2)} = 1 - \Phi(\tilde{z}_i^{(2)}) \), \( i = (i_1, \ldots, i_m) \),
where \( \tilde{z}_i^{(2)} \) denotes the standardized mean from the second stage
data for the hypothesis \( i \). The selection boundary \( \gamma_1 \) is specified in
the planning phase, whereas the rejection boundary \( \gamma_2 \) of the
second stage is chosen such that the FDR from the second stage equals 0.05.
Thus, asymptotically

\[
\alpha = \frac{\pi_0 \gamma_1}{\pi_0 \gamma_1 + (1 - \pi_0)(1 - \beta_2(\gamma_2))/(1 - \beta_1(\gamma_1))}
\]

Again, an optimization is carried out, and it can be seen that the
power of a pilot study (experiments ignoring the first stage data) is
only little smaller than of a two-stage design, but much larger than
the power of a single-stage design (compare Table 4 and Table 1).

Finally, we investigate the robustness of the optimal two-stage and
pilot study designs if the actual parameters deviate from the assump-
tions made in the planning phase. To demonstrate the differences,
this time we look at the scenario with a smaller overall sample size,
\( N = 4m_1 \). Now, for the choice of the value \( r \), only two scenarios are
possible for the two-stage design to get reasonable integer sample
sizes for the first stage, \( r = 0.5 \) with \( n_1 = 2 \), and \( r = 0.75 \) with
\( n_1 = 3 \). We choose \( r = 0.75 \) and \( \gamma_1 = 0.1 \), which is nearly optimal
for \( \pi_0 = 0.99 \) and \( \Delta_1/\sigma = 1.2 \). We calculate the asymptotic power
for the scenarios \( \pi_0 = 0.98 \), \( \pi_0 = 0.99 \) and \( \pi_0 = 0.995 \),
and \( \Delta_1/\sigma = 0.8, \Delta_1/\sigma = 1 \) and \( \Delta_1/\sigma = 1.2 \). As shown in Table 5,
the pilot study has a lower power than the two-stage study, since it does
not use the first stage data in the final test statistics. The relative
difference in power increases as \( \Delta_1/\sigma \) deviates from the parameter
value \( \Delta_1/\sigma = 1.2 \).

Obviously, the method of the pilot design can be directly applied to the
unknown variance case by using the single-stage \( P \)-values of the
\( t \)-test applied to the second stage data. Also the two-sided test
can be easily treated by taking the corresponding two-sided second
stage \( P \)-values.

Storey et al. (2004) recommend to apply a modified estimate of
the FDR if the number of hypotheses to be tested is small. This might
be the case in the pilot design if only a few hypotheses are selected

3775

Two-stage designs

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/21/19/3771/210591 by guest on 11 March 2019
proportions \( \pi \) with known variance, we derive optimal designs in terms of power.

In the first stage, promising hypotheses are selected over the hypotheses in a single-stage design, a two-stage design may be preferable considering the large second stage costs. In such situations it may be checked whether simple single-stage designs may be applied. Building up the devices for the second stage when the proportion of selected null hypotheses will be small. This is the reason why it is difficult to find real data in this area, where the two step procedure could be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

7 CONCLUSIONS

This manuscript deals with situations where a large number of hypotheses are investigated applying sample sizes that are constrained by costs. Such situations arise, e.g. in gene association studies, where a large number of markers are tested. Instead of distributing the sample sizes over the hypotheses in a single-stage design, a two-stage design is considered. In the first stage, promising hypotheses are selected for further investigation at the second stage. A multiple testing procedure based on data from both stages is proposed to control the FDR.

Assuming independently and normally distributed observations with known variance, we derive optimal designs in terms of power.

The optimal designs depend on the total sample size, the number of hypotheses to be investigated, the a-priori assumption on the proportion of true null hypotheses and a common effect size among the alternatives. The two-stage procedure shows striking superiority in terms of power as compared with the corresponding single-stage design, where the total number of observations is equally distributed among the hypotheses. The performance of the procedure does not substantially decrease if optimal designs are used based on wrong a-priori assumptions on the proportion of true hypotheses and the effect size under the alternative. The procedure also controls the FDR if the effect sizes under the alternative are not the same for all hypotheses but distributed according to some probability distribution.

Even if the test statistics are moderately correlated across hypotheses, the procedure provides satisfactory control of the FDR. Note that the sequential \( P \)-values can be used also for the control of the FWE rate applying Bonferroni adjusted critical boundaries. If observations across hypotheses are independent we proved that also the sequential \( P \)-values corresponding to the true null hypotheses are independent. Thus, all procedures based on independent univariate \( P \)-values can be applied.

Even the pilot design, where the first stage data is used only to select promising hypotheses which are tested at the second stage using only second stage data, leads to a strong improvement in power compared with the single-stage design. If, due to misspecifications in the planning phase, non-optimal designs are applied, the two-stage design appears to be more robust in terms of power than the pilot design. However, the pilot design is simpler and can be easily applied in all types of testing situations. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

7 CONCLUSIONS

This manuscript deals with situations where a large number of hypotheses are investigated applying sample sizes that are constrained by costs. Such situations arise, e.g. in gene association studies, where a large number of markers are tested. Instead of distributing the sample sizes over the hypotheses in a single-stage design, a two-stage design is considered. In the first stage, promising hypotheses are selected for further investigation at the second stage. A multiple testing procedure based on data from both stages is proposed to control the FDR.

Assuming independently and normally distributed observations with known variance, we derive optimal designs in terms of power.

The optimal designs depend on the total sample size, the number of hypotheses to be investigated, the a-priori assumption on the proportion of true null hypotheses and a common effect size among the alternatives. The two-stage procedure shows striking superiority in terms of power as compared with the corresponding single-stage design, where the total number of observations is equally distributed among the hypotheses. The performance of the procedure does not substantially decrease if optimal designs are used based on wrong a-priori assumptions on the proportion of true hypotheses and the effect size under the alternative. The procedure also controls the FDR if the effect sizes under the alternative are not the same for all hypotheses but distributed according to some probability distribution.

Even if the test statistics are moderately correlated across hypotheses, the procedure provides satisfactory control of the FDR. Note that the sequential \( P \)-values can be used also for the control of the FWE rate applying Bonferroni adjusted critical boundaries. If observations across hypotheses are independent we proved that also the sequential \( P \)-values corresponding to the true null hypotheses are independent. Thus, all procedures based on independent univariate \( P \)-values can be applied.

Even the pilot design, where the first stage data is used only to select promising hypotheses which are tested at the second stage using only second stage data, leads to a strong improvement in power compared with the single-stage design. If, due to misspecifications in the planning phase, non-optimal designs are applied, the two-stage design appears to be more robust in terms of power than the pilot design. However, the pilot design is simpler and can be easily applied in all types of testing situations. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.

Using two-stage designs will generally change the data structure of experiments: very few observations for very many questions, large sample sizes for few questions selected after screening. The tendency is to spend a major part of the observations for screening purposes. Even different measurement procedures can be applied to the reduced set of hypotheses at the second stage.
REFERENCES


