Sample Size Estimation while Controlling False Discovery Rate for Microarray Experiments Using the ssize.fdr Package

by Megan Orr and Peng Liu

Introduction

Microarray experiments are becoming more and more popular and critical in many biological disciplines. As in any statistical experiment, appropriate experimental design is essential for reliable statistical inference, and sample size has a crucial role in experimental design. Because microarray experiments are rather costly, it is important to have an adequate sample size that will achieve a desired power without wasting resources.

For a given microarray data set, thousands of hypotheses, one for each gene, are simultaneously tested. Storey and Tibshirani (2003) argue that controlling false discovery rate (FDR) is more reasonable and more powerful than controlling family-wise error rate (FWER) in genomic data. However, the most common procedure used to calculate sample size involves controlling FWER, not FDR.

Liu and Hwang (2007) describe a method for a quick sample size calculation for microarray experiments while controlling FDR. In this paper, we introduce the R package ssize.fdr which implements the method proposed by Liu and Hwang (2007). This package can be applied for designing one-sample, two-sample, or multi-sample experiments. The practitioner defines the desired power, the level of FDR to control, the proportion of non-differentially expressed genes, as well as effect size and variance. More specifically, the effect size and variance can be set as fixed or random quantities coming from appropriate distributions. Based on user-defined parameters, this package creates plots of power vs. sample size. These plots allow for visualization of trade-offs between power and sample size, in addition to the changes between power and sample size when the user-defined quantities are modified.

For more in-depth details and evaluation of this sample size calculation method, please refer to Liu and Hwang (2007).

Method

For a given microarray experiment, for each gene, let \( H = 0 \) if the null hypothesis is true and \( H = 1 \) if the alternative hypothesis is true. In a microarray experiment, \( H = 1 \) represents differential expression for a gene, whereas \( H = 0 \) represents no differential expression. As in Storey and Tibshirani (2003), we assume each test is Bernoulli distributed with the probability \( \Pr(H = 0) = \pi_0 \), where \( \pi_0 \) is the proportion of non-differentially expressed genes. Liu and Hwang (2007) derived that the following must hold to control FDR at the level of \( \alpha \):

\[
\frac{\alpha}{1 - \alpha} \cdot \frac{1 - \pi_0}{\pi_0} \geq \frac{\Pr(T \in \Gamma | H = 0)}{\Pr(T \in \Gamma | H = 1)}
\]

(1)

where \( T \) is the test statistic and \( \Gamma \) is the rejection region of the test. For each proposed hypothesis (test) with given user-defined quantities, the sample size is calculated using the following steps:

1. Solve for \( \Gamma \) using (1) for each sample size.
2. Calculate the power corresponding to \( \Gamma \) with appropriate formula for \( \Pr(T \in \Gamma | H = 1) \).
3. Determine sample size based on desired power.

For specific experimental designs, the numerator and denominator of the right hand side of (1) is replaced by corresponding functions that calculate the type I error and power of the test, respectively.

Functions

The package ssize.fdr has six functions which includes three new functions that have recently been developed as well as three functions translated from the previous available Matlab codes. The six functions and their descriptions are listed below:

```r
ssize.oneSamp(delta, sigma, fdr = 0.05, power = 0.8, pi0 = 0.95, maxN = 35, side = "two-sided", cex.title=1.15, cex.legend=1)

ssize.oneSampVary(deltaMean, deltaSE, a, b, fdr = 0.05, power = 0.8, pi0 = 0.95, maxN = 35, side = "two-sided", cex.title=1.15, cex.legend=1)

ssize.twoSamp(delta, sigma, fdr = 0.05, power = 0.8, pi0 = 0.95, maxN = 35, side = "two-sided", cex.title=1.15, cex.legend=1)

ssize.twoSampVary(deltaMean, deltaSE, a, b, fdr = 0.05, power = 0.8, pi0 = 0.95, maxN = 35, side = "two-sided", cex.title=1.15, cex.legend=1)
```
A study could be used. In this case, we recommend available, a guess based on a biological system un-
obtained from a pilot study. If a pilot study is not
among others quantities. The value of $\pi$

$\text{ssize.F}(X, \text{beta, L = NULL, dn, sigma,} $
\text{fdr = 0.05, power = 0.8, pi0 = 0.95,}$
\text{maxN = 35, cex.title=1.15,}$
\text{cex.legend=}1)$

$\text{ssize.FVary}(X, \text{beta, L = NULL, dn, a, b,}$
\text{fdr = 0.05, power = 0.8, pi0 = 0.95,}$
\text{maxN = 35, cex.title=1.15,}$
\text{cex.legend=}1)$

$\text{ssize.oneSamp}$ and $\text{ssize.twoSamp}$ compute ap-
propriate sample sizes for one- and two-sample mi-
croarray experiments, respectively, for fixed effect
size (delta) and standard deviation (sigma). For one-
sample designs, the effect size is defined as the dif-
fERENCE in the true mean expression level and its pro-
posed value under the null hypothesis for each gene.
For two-sample designs, the effect size is defined as
the difference in mean expression levels between
treatment groups for each gene. In the two-sample
case, sigma is the pooled standard deviation of treat-
ment expression levels.

$\text{ssize.oneSampVary}$ and $\text{ssize.twoSampVary}$
compute appropriate sample sizes for one- and
two-sample microarray experiments, respectively,
in which effect sizes and standard deviations vary
among genes. Effect sizes among genes are assumed
to follow a normal distribution with mean deltaMean
and standard deviation deltaSE, while variances (the
square of the standard deviations) among genes are
assumed to follow an Inverse Gamma distribution
with shape parameter $a$ and scale parameter $b$.

$\text{ssize.F}$ computes appropriate sample sizes for
multi-sample microarray experiments. Additional
inputs include the design matrix ($X$), the parameter
vector ($\text{beta}$), the coefficient matrix for linear con-
trasts of interest ($L$), a function of $n$ for the degrees
of freedom of the experimental design ($dn$), and the
pooled standard deviation of treatment expression
levels ($\text{sigma}$), assumed to be identical for all genes.

$\text{ssize.FVary}$ computes appropriate sample sizes for
multi-sample microarray experiments in which the
parameter vector is fixed for all genes, but the
variances are assumed to vary among genes and fol-
low an Inverse Gamma distribution with shape pa-
ter parameter $a$ and scale parameter $b$.

All functions contain a default value for $\pi_0$ ($\text{pi0}$),
among others quantities. The value of $\pi_0$ can be
obtained from a pilot study. If a pilot study is not
available, a guess based on a biological system un-
der study could be used. In this case, we recommend
using a conservative guess (bigger values for $\pi_0$) so
that the desired power will still be achieved. The in-
put $\pi_0$ can be a vector, in which case separate calcu-
lations are performed for each element of the vector.
This allows one to assess the changes of power due
to the changes in the $\pi_0$ estimate(s).

For functions that assume variances follow an In-
verse Gamma distribution, it is important to note
that if $\frac{1}{\sigma^2} \sim \text{Gamma}(a, \beta)$ with mean $a\beta$, version 1.1 of
the $\text{ssize.fdr}$ package uses the parameterization that
$\sigma^2 \sim \text{Inverse Gamma}(a,\beta^{-1})$.

Each function outputs the following results: a
plot of power vs. sample size, the smallest sample
size that achieves the desired power, a table of cal-
culated powers for each sample size, and a table of
calculated critical values for each sample size.

**Examples**

Unless otherwise noted, all data sets analyzed in this
section can be downloaded at http://bioinf.wehi.

**Using ssize.oneSamp**

and **ssize.oneSampVary**

To illustrate the use of the functions $\text{ssize.oneSamp}$
and $\text{ssize.oneSampVary}$, a data example from Smyth
(2004) will be used. In this experiment, zebrafish
are used to study early development in vertebrates.
Swirl is a point mutation in the BMP2 gene in ze-
brafish that affects the dorsal/ventral body axis. The
goal of this experiment is to identify gene expres-
sion levels that differ between zebrafish with the
swirl mutation and wild-type zebrafish. This data
set (Swirl) includes 8448 gene expression levels from
both types of the organism. The experimental de-
sign includes two sets of dye-swap pairs. For more
details on the microarray technology used and the
experimental design of this experiment, see Smyth
(2004). The **limma** package performs both nor-
malization and analysis of microarray data sets. See
Section 11.1.1 of Smyth et al. (2008) for complete anal-
ysis of the Swirl data set, along with the R code for this
analysis.

After normalizing the data, the **limma** package is
applied to test if genes are differentially expressed
between groups, in this case genotypes (Smyth et al.,
2008). This is done by determining if the $\log_2$ ratio
of mean gene expressions between swirl mutant ze-
brafish and the wild type zebrafish are signifi-
cantly different from zero or not. We then apply the
qvalue function (in the qvalue package) to the vector of p-
values obtained from the analysis, and $\pi_0$ is esti-
mated to be 0.63.

Now suppose we are interested in performing a
similar experiment and want to find an appropri-
ate sample size to use. We want to be able to de-
tect a two-fold change in expression levels between
the two genotypes, corresponding to a difference in
$\log_2$ expressions of one. We find both up-regulated
and down-regulated genes to be interesting, thus our
tests will be two-sided. We also want to obtain 80% power while controlling FDR at 5%. We will use the estimated value of $\pi_0$ (0.63) from the available data
in addition to other values for \( \pi_0 \) to estimate sample size.

**The ssize.oneSamp function:** To illustrate the use of the `ssize.oneSamp` function, we need to assign a common value for the standard deviations of all genes, so we choose the 90\(^{th}\) percentile of the sample standard deviations from the swirl experiment. This value is 0.44. It is important to note that power calculations will be overestimated for genes with higher standard deviation than this. Similarly, power will be conservative for genes with smaller standard deviation than 0.44. To calculate appropriate sample sizes using `ssize.oneSamp`, the following code is used.

```r
> os <- ssize.oneSamp(delta=1, sigma=0.44, 
                      fdr=0.05, power=0.8, 
                      pi0=c(0.5, 0.63, 0.75, 0.9), maxN=10)
```

This code produces the graph shown in Figure 1

![Power vs. sample size with fdr=0.05 and Δ/σ = 2.2727](image)

**Figure 1:** Sample size vs. power for one-sample two-sided t-test with effect size of one, standard deviation of 0.44, and various proportions of non-differentially expressed genes with FDR controlled at 5%.

We see that appropriate samples sizes are 4, 5, 6, and 7 for the above experiment at \( \pi_0 \) values of 0.50, 0.63, 0.75, and 0.90. Notice that although a sample size of five was calculated to be adequate for \( \pi_0 = 0.63 \), we are performing a dye-swap experiment, so it is better to have an even number of slides. Because of this, we recommend six slides for a dye-swap experiment with two pairs. Sample size information, along with the calculated powers and critical values can be obtained using the following code.

```r
> os$sssize
> os$spower
> os$scrit.vals
```

**The ssize.oneSampVary function:** Now suppose we are interested in representing the standard deviations of the log ratios more precisely. This can be done using a Bayesian approach. Smyth (2004) models the variances of normalized expression levels as

\[
\frac{1}{\sigma_s^2} \sim \frac{1}{d_0s_0^2}\chi^2_{d_0}
\]

where \( d_0 \) and \( s_0^2 \) are hyperparameters and can be estimated based on observed residual variances. Performing a simple transformation, we see that (2) is equivalent to

\[
\sigma_s^2 \sim \text{Inverse Gamma} \left( \frac{d_0}{2}, \frac{d_0s_0^2}{2} \right)
\]

where the expected value of \( \sigma_s^2 \) is \( d_0s_0^2/(d_0 - 2) \). Using the `lmFit` and `eBayes` functions in the `limma` package, \( d_0 \) and \( s_0^2 \) can be calculated from any microarray data set. For the Swirl data set, these values are calculated to be \( d_0 = 4.17 \) and \( s_0^2 = 0.051 \) (Smyth, 2004). From (3), we see that the variances of the log ratios follow an Inverse Gamma distribution with shape parameter 2.09 and scale parameter 0.106. From this information, we find appropriate sample sizes for a future experiment using the `ssize.oneSampVary` function as follows.

```r
> osv <- ssize.oneSampVary(deltaMean=1, 
                          deltaSE=0, a=2.09, b=0.106, 
                          fdr=0.05, power=0.8, 
                          pi0=c(0.5, 0.63, 0.75, 0.9), maxN=15)
```

Figure 2 shows the resulting plot of average power versus sample size for each proportion of non-differentially expressed genes that was input. Note that we are simply interested in finding genes that have a two-fold change in expression levels, so we want the effect size to be held constant at one. This is done by setting the mean of the effect sizes \( (\text{deltaMean}) \) at the desired value with standard deviation \( (\text{deltaSE}) \) equal to zero.

From Figure 2, we see that the appropriate sample sizes are 3, 4, 4, and 5, respectively, to achieve 80% average power with a FDR of 5%. These values are smaller than those in the previous example. In order to obtain appropriate sample size information more directly, along with the calculated powers and critical values, use the following code.

```r
> osv$sssize
> osv$spower
> osv$scrit.vals
```
The goal of this study is to determine how the absence of the ApoAI gene affects other genes in the liver. In order to do this, gene expression levels of ApoAI knockout mice and control mice were compared using a common reference design. See Smyth (2004) for a more in-depth description of the study. This is an example from Smyth (2004) will be used. The ApoAI gene plays an important role in high density lipoprotein (HDL) metabolism, as mice with this gene knocked out have very low HDL levels. Using the ApoAI data. From (3), it follows that the variance of the gene expression levels follow an Inverse Gamma distribution with shape parameter 1.94 and scale parameter 0.106.

Figure 2: Sample size vs. power for one-sample two-sided t-test with effect size of one and various proportions of non-differentially expressed genes with FDR controlled at 5%. Variances are assumed to follow an Inverse Gamma(2.09, 0.106) distribution.

Using ssize.twoSamp and ssize.twoSampVary

In order to illustrate the uses of the ssize.twoSamp and ssize.twoSampVary functions, another data example from Smyth (2004) will be used. The ApoAI gene plays an important role in high density lipoprotein (HDL) metabolism, as mice with this gene knocked out have very low HDL levels. The goal of this study is to determine how the absence of the ApoAI gene affects other genes in the liver. In order to do this, gene expression levels of ApoAI knockout mice and control mice were compared using a common reference design. See Smyth (2004) for a more in-depth description of the study. This is an example of a two sample microarray experiment. For full analysis of this data sets using the limma package, see Section 11.4 of Smyth et al. (2008).

Similar to the previous example, the limma package is applied to analyze the data. From the resulting p-values, we use the qvalue function to estimate the value of \( \pi_0 \) as 0.70.

The ssize.twoSamp function: To illustrate the use of the ssize.twoSamp function, we can choose a common value to represent the standard deviation for all genes. Similar to the ssize.oneSamp example, we can use the 90th percentile of the gene residual standard deviations. In this case, this value is 0.42. Again assume we are interested in genes showing a two-fold change, but this time we are only concerned with up-regulated genes, so we will be performing a one-sided upper tail t-test for each gene. We also want to achieve 80% power while controlling FDR at 5%. The ssize.twoSamp function can be used to calculate an appropriate sample size for a similar experiment with the following code. This code also produces Figure 3.

```r
> ts <- ssize.twoSamp(delta=1, sigma=0.42, fdr=0.05, power=0.8, pi0=c(0.6, 0.7, 0.8, 0.9), side="upper", maxN=15)
```

Figure 3: Sample size vs. power for two-sample one-sided upper t-test with effect size of one, standard deviation of 0.42, and various proportions of non-differentially expressed genes with FDR controlled at 5%.

From Figure 3, we see that appropriate sample sizes are 4, 5, 6, and 7 for each genotype group for \( \pi_0 \) values of 0.6, 0.7, 0.8, and 0.9, respectively. Calculated critical values and power along with the sample size estimates can be obtained with the following code.

```r
> ts$ssize
> ts$power
> ts$crit.vals
```

The ssize.twoSampVary function: As in the ssize.oneSampVary example, the limma package calculates \( \Delta_0 \) to be 3.88 and \( \Delta_i \) to be 0.05 as in (2) and (3) using the ApoAI data. From (3), it follows that the variances of the gene expression levels follow an Inverse Gamma distribution with shape parameter 1.94 and scale parameter 0.10. With these values, we can use the ssize.twoSampVary function to calculate appropriate sample sizes for experiments for detecting a fold change of one in up-regulated genes. We also want our results to have at least 80% average power with FDR controlled at 5%.

```r
> tsv <- ssize.twoSampVary(deltaMean=1, deltaSE=0, a=1.94, b=0.10,
```

Figure 2: Sample size vs. power for one-sample two-sided t-test with effect size of one and various proportions of non-differentially expressed genes with FDR controlled at 5%. Variances are assumed to follow an Inverse Gamma(2.09, 0.106) distribution.
µ represents the mean of normalized gene expression levels for cells with no estrogen at 10 hours, τ represents the effect of time in the absence of estrogen, α represents the change for cells with estrogen at 10 hours, and γ represents the change for cells with estrogen at 48 hours Smyth et al. (2008).

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>no estrogen, 10 hours</td>
<td>µ</td>
</tr>
<tr>
<td>estrogen, 10 hours</td>
<td>µ + τ</td>
</tr>
<tr>
<td>no estrogen, 48 hours</td>
<td>µ + α</td>
</tr>
<tr>
<td>estrogen, 48 hours</td>
<td>µ + τ + γ</td>
</tr>
</tbody>
</table>

One design matrix that corresponds to this experiment and the means in the table above is

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

with parameter vector $\beta = [\mu, \tau, \alpha, \gamma]$. Notice that for this design, there are four parameters to estimate and four slides for each sample, thus the degrees of freedom are $4n - 4$. Because we are only interested in the effect of estrogen and how this changes over time, we only care about the parameters $\alpha$ and $\gamma$. Thus, we let $L$ be the matrix of the linear contrasts of interest or

$$L = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

so that $L'\beta = [\alpha, \gamma]$.

In order to calculate samples sizes for a similar experiment that will result in tests with 80% power and FDR controlled at 5%, we must have a reasonable estimate for $\pi_0$. Using the limFit and eBayes functions of the limma package in conjunction with qvalue, we obtain an estimate value for $\pi_0$ of 0.944. Similar to the other examples, we will include this value along with other similar values of $\pi_0$ to see how power varies.

The ssize.F function: As in all of the previous examples, we find the 90th percentile of the sample residual standard deviations to be 0.29 (or we can find any other percentile that is preferred). Also, let the true parameter vector be $\beta = [12, 1, 1, 0.5]$. Note that the values of the $\mu$ and $\tau$ parameters do not affect any sample size calculations because we are only interested in $\alpha$ and $\gamma$, as represented by $L'\beta$. Thus, the following code will calculate the sample size required to achieve 80% power while controlling the FDR at 5%.

```r
> X <- matrix(c(1,0,0,0,0,1,0,1,0,0,1,0,1,1,1,0,0,1,0,1,0,1,1,1,0,1,1,1,0,0,1,0,1,1,1,0,0,1), nrow=4, byrow=TRUE)  
> L <- matrix(c(0,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1), nrow=4, byrow=TRUE)  
> B <- c(12, 1, 1, 0.5)
```
Figure 5: Sample size vs. power for F-test with parameter vector $\beta = [12, 1, 1, 0.5]$, common gene expression standard deviation of 0.29, and various proportions of non-differentially expressed genes with FDR controlled at 5%.

From Figure 5, we see that appropriate sample sizes for this experiment are 4, 5, 6, for $\pi_0$ values of 0.9, 0.944, 0.975, and 0.995, respectively. It is important to note that in this experiment, one sample includes a set of four gene chips (one for each treatment). So a sample size of 4 would require a total of 16 chips, with 4 for each of the four treatment groups.

For other useful information, use the following code.

```r
> dn <- function(n){4*n - 4}
> fs <- ssize.F(X=X, beta=B, L=L, dn=dn, sigma=0.29, pi0=c(0.9, 0.944, 0.975, 0.995), maxN=12)
```

The `ssize.Fvary` function: Using the limma package, $d_0$ and $s_0^2$ are calculated as 4.48 and 0.022, where $d_0$ and $s_0^2$ are as in (2). As shown in (3), this corresponds to inverse gamma parameters for $c_0^2$ of 2.24 and 0.049. Using the same parameter vector as in the previous example, sample sizes for future experiments can be calculated with the following commands, where we wish to control the FDR at 5% and achieve an average test power of 80%.

```r
> fs <- ssize.Fvary(X=X, beta=B, L=L, dn=dn, a=2.24, b=0.049, pi0=c(0.9, 0.944, 0.975, 0.995), maxN=12)
```

Figure 6: Sample size vs. power for F-test with parameter vector $\beta = [12, 1, 1, 0.5]$, gene expression variances following an Inverse Gamma(2.24,0.049), and various proportions of non-differentially expressed genes with FDR controlled at 5%.

Figure 6 suggests that appropriate sample sizes for this experiment are three for the smallest three $\pi_0$ values and four for $\pi_0=0.995$. For calculated sample sizes, critical values, and powers, use the following code.

```r
> fsv$ssize
> fsv$power
> fsv$crit.vals
```

Modifications

Functions `ssize.oneSampVary`, `ssize.twoSampVary`, and `ssize.Fvary` calculate sample sizes using the assumption that at least one parameter is random. Currently, these assumptions are that effect sizes follow Normal distributions and variances follow Inverse Gamma distributions. If users desire to assume that parameters follow other distributions, the code can be modified accordingly.

Acknowledgement

This material is based upon work partially supported by the National Science Foundation under Grant Number 0714978.

Bibliography

P. Liu and J. T. G. Hwang. Quick calculation for sample size while controlling false discovery rate with


Megan Orr
Department of Statistics & Statistical Laboratory
Iowa State University, USA
meganorr@iastate.edu

Peng Liu
Department of Statistics & Statistical Laboratory
Iowa State University, USA
pliu@iastate.edu