Sample Size and Power Calculation in Microarray Studies Using the sizepower package.

Weiliang Qiu
email: weiliang.qiu@gmail.com
Mei-Ling Ting Lee
email: meilinglee@sph.osu.edu
George Alex Whitmore
email: george.whitmore@mcgill.ca

April 16, 2015

1 Introduction

The process of obtaining biological samples is often expensive, involved, and time consuming. Thus, the issue of the appropriate sample size is important in planning a study. If the sample size is too large, we waste resources. If the sample size is too small, we cannot draw inferences with the desired precision. Thus, we need to calculate the sample size of a study before proceeding in order to determine the best trade-off between precision and resource use.

The calculation of the sample size is closely related to that of power. One goal of microarray studies is to find a subset of genes that are differentially expressed across experimental conditions. The key question is the strength of the claim that a given gene is differentially expressed. In other words, what is the power of the test to determine if a gene is differentially expressed to a specified degree?

The R package, sizepower, is used to calculate sample size and power in the planning stage of a microarray study. It helps the user to determine how many samples are needed to achieve a specified power for a test of whether a gene is differentially expressed or, in reverse, to determine the power of a given sample size.

This R package provides two functions for sample-size calculations for two types of experimental designs (a completely randomized treatment-control design and a matched-pairs design) and three functions for power calculation for four types of experimental designs (a completely randomized treatment-control design, a matched-pairs design, a multiple-treatment design having an isolated treatment effect, and a randomized block design).

2 Experiment Designs

The following discussion of sample size planning for microarray studies assumes that the investigator has already chosen a particular microarray technology and selected the measure of gene expression. The statistical testing proceeds after the application of appropriate background correction, normalization and mathematical transformations. Thus, expression levels in the following discussion of testing refer to data that have been through these preprocessing steps.

2.1 Completely randomized treatment-control designs

In this design, we consider two groups of biological samples: a treatment group and a control group. We are interested in testing if the mean expression level for a given gene is the same for the two groups. The hypotheses of interest for any specific gene are

$$H_0 : \theta_t = \theta_c \quad \text{vs} \quad H_1 : \theta_t \neq \theta_c, \tag{1}$$

where θ_t and θ_c are the mean expression levels of the treatment and control groups, respectively, for the given gene. When we proceed to consider the required sample size or power level, we will let $\mu_1 = \theta_t - \theta_c$ denote the specific difference in means postulated in the alternative hypothesis H_1.

All statistical observations in the design are assumed to be mutually independent. In addition, after appropriate mathematical transformation (such as a log-transformation), all observations in the treatment group are assumed to be drawn from the normal distribution $\mathrm{N}(\theta_t,\ \sigma^2)$, while samples in the control group are assumed to be drawn from the normal distribution $\mathrm{N}(\theta_c,\ \sigma^2)$, where the two distributions share a common variance σ^2. To simplify the presentation, we only consider the case where each group has the same number of observations n.

For example, in a microarray toxicity study, we randomly assign $2n$ mice in equal numbers to treatment and control groups. The n mice in the treatment group will be exposed to a toxin and the n mice in the control group will not be exposed to the toxin. We are interested in detecting if any mouse gene on the microarray will be differentially expressed between the treatment and control groups.

2.2 Matched-pairs designs

Like the completely randomized treatment-control design, the matched-pairs design consists of two groups of biological samples and we are interested in whether the mean expression levels for genes from the two groups are different. In the matched-pairs design, however, the observations from the two groups are not independent. Instead, each treatment sample is paired with one control sample, creating n pairs of matched (correlated) observations. Within a matched pair, the two observations are dependent. The matched pairs themselves are independent. We still assume that observations in the treatment

and control groups are normally distributed with means θ_t and θ_c, respectively, and with common variance σ^2.

For example, in a microarray liposarcoma study, we have n patients. From each patient, we obtain one sample of liposarcoma tissue and one sample of normal fat tissue, giving a total of n pairs of tissue samples. Within each pair, the two tissue samples share some features because they are taken from the same patient. We are interested in detecting genes on the microarray that are differentially expressed in the two types of tissue (liposarcoma and normal fat tissues). In other words, for any single gene, we would like to test if the mean difference in expression levels of the two types of tissue is equal to zero.

2.3 Multiple-treatment designs having an isolated treatment effect

This design is a generalization of the completely randomized treatment-control designs to multiple treatments. Suppose there are T treatments, designated $t = 1, \ldots, T$. We first randomly divide nT biological samples into T groups, each with n samples. Then the t-th group is assigned the t-th treatment. The hypotheses are:

$$\begin{aligned} H_0: &\quad \theta_1 = \cdots = \theta_T \\ H_1: &\quad \text{One treatment mean differs from the other } T-1 \text{ means,} \end{aligned} \tag{2}$$

where θ_t, $t = 1, \ldots, T$, are the gene expression means of the T treatment groups. The term *isolated effect design* refers to the fact that, under the alternative hypothesis, all treatments are assumed to have the same mean level of expression except for one treatment that has a different mean level. When we proceed to consider the required sample size or power level, we will let μ_1 denote the specific difference between the mean of the isolated treatment and the common mean of the remaining $T-1$ treatments under H_1.

We assume that all observations for the nT samples are mutually independent and that observations from the t-th treatment group are normally distributed with mean θ_t and variance σ^2. Thus, we assume the treatment groups have a common variance. As already seen, we also assume that all treatment groups have the same number of observations.

2.4 Randomized block designs having an isolated treatment effect

This design is a generalization of matched-pairs designs to multiple treatments. The benefits of matching are achieved when biological samples within the same block are more homogeneous than samples in different blocks. For example, mice in the same litter tend to be more similar than mice in different litters. A total of n blocks are constructed with each block having T biological samples. The T treatments are randomly applied

to the samples within each block. For each gene, we are interested in testing if all treatments have the same mean level of expression except for one treatment that has a different mean level.

Within a block, the expression levels of a gene are assumed to be dependent. Between blocks, the expression levels of a gene are assumed to be independent. We assume that the expression level of a gene in the t-th treatment group is normally distributed with mean θ_t and variance σ^2, $i = 1 \ldots, T$.

3 Sample-size Calculation

In this section, we show the method of sample-size calculation for completely randomized treatment-control designs and matched-pairs designs.

Denote the total number of genes in a microarray study by G. Next, let G_0 be the true number of genes that are not differentially expressed and R_0 be the number of genes that are not differentially expressed but are falsely declared by the test procedure to be differentially expressed, i.e., R_0 is the number of false positives. The count G_0 is some fixed but unknown number. Prior to applying the test procedure to the data, R_0 is an unknown and random count.

The probability α_0 of a type I error for any single gene g among the G_0 genes that are not differentially expressed is given by

$$\alpha_0 = \frac{E(R_0)}{G_0}, \tag{3}$$

where $E(R_0)$ is the expected number of false positives from applying the test procedure to all of the G_0 genes.

Recall that we let $\mu_1 = \theta_t - \theta_c$ denote the difference in means postulated in the alternative hypothesis H_1. For specified values of $E(R_0)$ and G_0, we can calculate the requisite sample size to achieve a specified power for completely randomized treatment-control designs and matched-pairs designs as follows:

$$n = \left(\frac{z_a + z_b}{|\mu_1|/\sigma_d} \right)^2, \tag{4}$$

where n is the sample size for each group, $a = 1 - \alpha_0/2$, b is the power of the test, z_a and z_b are the lower a and b percentiles of the standard normal distribution, and σ_d is the standard deviation of the difference in expression between treatment and control samples, as defined below.

For a completely randomized treatment-control design, σ_d^2 is the variance of the difference in expression between randomly chosen treatment and control samples. For a matched-pairs design, σ_d^2 is the variance of the difference in expression between the treatment and control samples in a matched pair. In both cases,

$$\sigma_d^2 = 2\sigma^2, \tag{5}$$

where σ^2 is the variance of the random error component in an ANOVA model.

For a completely randomized treatment-control design, the ANOVA model for any single gene is

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}, i = 1, \ldots, n, j = 1, 2 \tag{6}$$

where y_{ij} is the expression level of the i-th biological sample in the j-th group, μ is the overall mean, τ_j is the j-th treatment effect, and ϵ_{ij} is the random error component and is normally distributed with mean 0 and variance σ^2. Note that $\theta_t = \mu + \tau_1$ and $\theta_c = \mu + \tau_2$ where $\tau_1 + \tau_2 = 0$.

For a matched-pairs design, the ANOVA model for any single gene is

$$y_{ij} = \mu + \tau_j + \beta_i + \epsilon_{ij}, i = 1, \ldots, n, j = 1, 2 \tag{7}$$

where y_{ij} is the expression level of the i-th biological sample in the j-th group, μ is the overall mean, τ_j is the j-th treatment effect, β_i is the i-th block effect, and ϵ_{ij} is the random error component and is normally distributed with mean 0 and variance σ^2. Note that $\theta_t = \mu + \tau_1$ and $\theta_c = \mu + \tau_2$.

To calculate the sample size in formula (4), we have to anticipate the values of G_0 and σ_d. The investigator must also specify the desired values of $E(R_0)$, μ_1 and the power b. Typically, a high percentage of the genes are not differentially expressed in a microarray study. In this case, the total number of genes G might be substituted for G_0 in the calculation $\alpha_0 = E(R_0)/G_0$, giving the approximation $\alpha_0 \simeq E(R_0)/G$.

We let α_F denote the probability that one or more false positives will be produced in testing the family of G_0 genes that are not differentially expressed. The probability of the type I error for an individual test α_0, the probability of the type I error for the family of tests α_F, and the false positive rate $E(R_0)$ are interrelated. Under the Šidák approach to type I error control, we have the following approximation (Lee, 2004, page 202):

$$E(R_0) = \alpha_0 G_0 \simeq -\ln(1 - \alpha_F). \tag{8}$$

Under the Bonferroni approach to type I error control, we have the following approximation (Lee, 2004, page 203):

$$E(R_0) = \alpha_0 G_0 \simeq \alpha_F. \tag{9}$$

A subject matter expert can estimate G_0. Similarly, a pilot study or closely related study can be used to estimate the value of σ^2 and, hence, σ_d^2. For example, we can calculate an ANOVA table for each gene from pilot study data and estimate σ^2. (In an ANOVA test, σ^2 is estimated by the mean square error [MSE]). We can then average these estimates to obtain a pooled estimate of σ^2.

The user may wish to explore the sensitivity of the sample size to a range of specifications for $E(R_0)$, μ_1 and the power b. For the power b, for example, the user may try a series of probabilities, such as 0.7, 0.8, 0.9 and 0.95, to get a series of sample sizes.

To remind the user that completely randomized treatment-control designs and matched-pairs designs are different, we provide one function for each design to implement the sample-size calculation.

4 Power Calculation

In this section, we present a unified methodology for computing power for all four designs, namely, completely randomized treatment-control designs, matched-pairs designs, multiple-treatment designs having an isolated treatment effect, and randomized block designs. In this methodology, we obtain the power value by solving the following equation system for (c, b):

$$\begin{aligned} \alpha_0 &= Pr(\chi^2 > c | H_0 \text{ true}) \\ b &= Pr(\chi^2 > c | H_1 \text{ true}). \end{aligned} \tag{10}$$

Here χ^2 is the test statistic. The value c is a cutoff that determines whether to reject H_0 or not. We reject H_0 if $\chi^2 > c$ and do not reject it otherwise. The value b is the power of the test and $\alpha_0 = E(R_0)/G_0$ is the probability of the type I error for an individual test.

For all four designs we discussed (i.e., completely randomized treatment-control designs, matched-pairs designs, multiple-treatment designs having an isolated treatment effect, and randomized block designs), the test statistic χ^2 is chi-square distributed with $T-1$ degrees of freedom under the null hypothesis H_0, where T is the number of treatments. Under H_1, the test statistic χ^2 has a non-central chi-square distribution with non-centrality parameter

$$\psi_1 = \frac{n(T-1)}{T} \left(\frac{|\mu_1|}{\sigma} \right)^2, \tag{11}$$

where the notation is that defined earlier. Recall, specifically, that μ_1 denotes the difference $\theta_t - \theta_c$ for the treatment-control designs and the magnitude of the isolated effect for the multiple treatments designs. The parameter σ^2 in each design is the variance of the ANOVA error term.

We obtain estimates of σ and G_0 by the same method discussed in the previous section. Also, the specifications for $|\mu_1|$, $E(R_0)$ and power level b are required from the investigator in the same manner as described previously.

We provide one function to calculate the power for completely randomized treatment-control design, one function for matched-pairs designs, and one function for multiple-treatment designs having an isolated treatment effect and randomized block designs.

5 Examples

To call the functions in the `R` package sizepower, we first need to load it into `R`:

Consider a completely randomized treatment-control design with an equal number of biological samples in the treatment and control groups. It is anticipated that $G_0 = 2000$ genes will not be differentially expressed. The mean number of false positives is to be controlled at $E(R_0) = 1$ and power is to be controlled at $|\mu_1| = 1.00$. From similar studies done previously, it is anticipated that $\sigma = 0.40$ (i.e., $\sigma_d = \sqrt{2}(0.40) = 0.566$). If

we wish the power level to be $b = 0.9$, we can use the following command to calculate the number of samples needed for each group:

```
> sampleSize.randomized(ER0=1, G0=2000, power=0.9, absMu1=1, sigmad=0.566)

$n
[1] 8

$d
[1] 1.766784
```

That is, 8 samples are needed for each group. This sample size is the smallest n that will provide the required power. The returned value d is the statistical distance between treatment and control conditions specified under H_1, i.e., $d = |\mu_1|/\sigma_d$.

If we want to see what the exact power is if $n = 8$, we can use the command:

```
> power.randomized(ER0=1, G0=2000, absMu1=1, sigmad=0.566, n=8)

$power
[1] 0.9352991

$psi1
[1] 24.97222
```

That is, the exact power is 0.935 and the non-centrality parameter of the associated non-central chi-square statistic is 24.97.

Consider a matched-pairs design. Suppose a pilot study shows that $\sigma = 0.35$ (i.e., $\sigma_d = \sqrt{2}(0.35) = 0.495$). We wish $E(R_0) = 1$, $|\mu_1| = 1.00$, and the power $b = 0.90$. We also expect that $G_0 = 2000$. Then we can use the command:

```
> sampleSize.matched(ER0=1, G0=2000, power=0.9, absMu1=1, sigmad=0.495)

$n
[1] 6

$d
[1] 2.020202
```

If we want to see what the exact power is if $n = 6$, we can use the command:

```
> power.matched(ER0=1, G0=2000, absMu1=1, sigmad=0.495, n=6)

$power
[1] 0.9289082

$psi1
[1] 24.4873
```

That is, the exact power is 0.929 and the non-centrality parameter is 24.487.

Consider a multiple-treatment design involving $T = 5$ treatments and $G_0 = 2000$ undifferentially expressed genes. Suppose we wish to control $E(R_0) = 1.0$ and to detect an isolated effect of $|\mu_1| = 1.00$ between one distinguished treatment and all other treatments. We anticipate an experimental error standard deviation of $\sigma = 0.40$. To calculate the power for $n = 6$, we can use the following command:

```
> power.multi(ER0=1, G0=2000, numTrt=5, absMu1=1, sigma=0.4, n=6)

$power
[1] 0.9049313

$psi1
[1] 30
```

That is, the power is 0.904 and the non-centrality parameter is 30.

Consider a randomized block design with $T = 6$ treatments, $n = 9$ blocks, and $G_0 = 5000$. We wish $E(R_0) = 1$ and $|\mu_1| = 0.9$. From a pilot study, we anticipate that $\sigma = 0.5$. Then the power is

```
> power.multi(ER0=1, G0=5000, numTrt=6, absMu1=0.9, sigma=0.5, n=9)

$power
[1] 0.662967

$psi1
[1] 24.3
```

References

[1] Lee, M.-L. T. *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers, 2004.

[2] Lee, M.-L. T. and Whitmore, G. A. Power and sample size for DNA microarray studies. *Statistics in Medicine*, 21:3543–3570, 2002.