

Sample Size Calculation Routine Library (version 2)

By Stan Pounds
September 18, 2008

Introduction

This is a guide on how to use the R routine library for implementing the Pounds and Cheng ([2005, *Bioinformatics*](#)) method to calculate the sample size for an experiment in which the final data analysis will consist of performing many one-way ANOVA analyses and using the false discovery rate (or similar metrics such as the conditional false discovery rate or positive false discovery rate) to account for multiple testing. One application is to provide statistical guidance for choosing the sample size for a microarray study that compares gene expression across $k \geq 2$ independent groups. Please note that version 2 rectifies some discrepancies in how the documentation of the original version defined the non-centrality parameter of the non-central F distribution. Version 2 uses the same definition of the non-centrality parameter throughout all documentation.

The routine library defines a number of functions. The end-user will be primarily interested in the functions `sampsize.oneway` and `bgtheta.oneway`. The function `sampsize.oneway` performs the sample size calculations and the function `bgtheta.oneway` uses F-statistics from background data to compute a value of the θ parameter to use in `sampsize.oneway`. Each of these two functions is described in greater detail below. Following their description, there are a couple of examples.

The Function `sampsize.oneway`

The function `sampsize.oneway` accepts the following arguments:

- (1) `tau` – a scalar giving the desired level of FDR control in the final analysis.
- (2) `delta` – a scalar giving the average statistical power, i.e. the desired proportion of true alternatives to be detected as significant.
- (3) `theta` – a vector giving the values of the theta parameter for each statistical test to be performed. The θ parameter is one-half of the ratio of the sum of squared differences between each group mean and an overall mean to within-group variance. See equations (18) and (19) of Pounds and Cheng ([2005, *Bioinformatics*](#)) for the detailed mathematical description. Note that the definition of theta differs from the definition of the non-centrality parameter in R. The function `bgtheta.oneway` can be used to compute a value of theta if there is background data from the same groups for which one-way ANOVA F-statistics can be computed.
- (4) `ngrps` – an integer specifying the number k of experimental groups
- (5) `BH` - a logical (TRUE/FALSE) indicating whether the original FDR control procedure (Benjamini and Hochberg 1995, J. Roy. Stat. Soc. B) will be used for the final analysis. The default is FALSE.
- (6) `n0` – the per-group sample size at which to begin iterations. The default is 2. The algorithm begins with this sample size and increments until the conditions

for tau and delta are satisfied or it reaches a maximum sample size, specified in the next argument.

- (7) `nmax` – the maximum sample size to be considered in the calculations. This allows the algorithm to stop if it becomes apparent that the sample size satisfying the requirements is unfeasibly large.

The function `sampsize.oneway` returns a list object with the following components:

- (1) `n` – the computed sample size
- (2) `alpha` – the computed p-value threshold that is anticipated to result in the desired FDR control and average statistical power.
- (3) `aFDR` – the anticipated false discovery rate at alpha given n
- (4) `apow` – the anticipated average power at alpha given n
- (5) `tau` – the desired level of FDR control specified by the user
- (6) `delta` – the desired average statistical power specified by the user
- (7) `OK` – a logical variable (TRUE/FALSE) indicating whether the requirements stated by the user were satisfied by n

The Function `bgtheta.oneway`

The function `bgtheta.oneway` accepts the following arguments:

- (1) `Fstats` – a vector of one-way ANOVA F-statistics computed from the background data
- (2) `df1` – the numerator degrees of freedom for the F-statistics in `Fstats`
- (3) `df2` – the denominator degrees of freedom for the F-statistics in `Fstats`
- (4) `fdr.adj` – a logical variable (TRUE/FALSE) indicating whether to use the spacings loess histogram procedure (SPLOSH; [Pounds and Cheng 2004, *Bioinformatics*](#)) to adjust estimates of theta for multiplicity. The default is TRUE.

If `fdr.adj` is set to TRUE, then the function `bgtheta.oneway` returns a list object with the following components:

- (1) `theta` – a vector of unadjusted estimates of theta
- (2) `fdr.theta` – a vector of the adjusted estimates of theta
- (3) `fdrres` – a list giving details of the results of the SPLOSH procedure, see the user's guide to SPLOSH for more information

If `fdr.adj` is set to FALSE, then the function `bgtheta.oneway` returns a vector of unadjusted theta estimates.

Examples

Suppose one wishes to compute the sample size required to have an average statistical power of 50% to detect differences between two groups while controlling the FDR to be

10% or less. A total of 10,000 statistical tests are to be performed. It is believed that the null hypothesis is true for 9,000 of these tests and that $\theta = 1$ for the remaining tests. Benjamini and Hochberg's (1995) procedure will be used for the final analysis. The largest feasible per-group sample size is 50. Here is an example of how to perform the calculation and view the results:

```
# Read in the routine library
source("SampSizeLibrary.ssc")

# Define the theta vector for the calculation
theta<-c(rep(0,9000),rep(1,1000))

# Call sampsize.oneway and store the result in "calc"
calc<-sampsize.oneway(tau=0.10,
                      delta=0.50,
                      theta=theta,
                      ngrps=2,
                      BH=TRUE,
                      n0=2,
                      nmax=50)

# Display the contents of "calc"
print(calc)
```

The results are shown below.

```
$n
[1] 6

$alpha
[1] 0.005147294

$aFDR
[1] 0.09421659

$apow
[1] 0.5

$tau
[1] 0.1

$delta
[1] 0.5

$OK
[1] TRUE
```

The OK component indicates that the stated requirements the computed sample size satisfies the stated statistical requirements. The n component indicates that a per-group sample size of 6 is needed. Based on the calculations, it is anticipated that Benjamini and Hochberg's (1995) procedure will choose $\alpha = 0.0051$ (or a larger value) as the p-value cut-off. At this value of α , we anticipate that no more than 9.4% of the significant

findings will be false discoveries and that 50% of the true alternatives will be declared significant.

If a pilot data set for the groups of interest is available, then apply one-way ANOVA to the data set to compute the F-statistics. Suppose these F-statistics are stored in a vector called `Fstats`, the numerator degrees of freedom stored in `df1`, the denominator degrees of freedom stored in `df2`. Then, the following code will compute and display the results.

```
# Use the background data to compute theta for the sample size calculation
# Compute FDR-adjusted theta estimates
# and store results in the object "bg.theta"
bg.theta<-bgtheta.oneway(Fstats,df1,df2,fdr.adj=TRUE)

# Extract the FDR-adjusted theta estimates
# and store then in a vector called "theta"
theta<-bg.theta$fdr.theta

# Call sampsize.oneway and store the result in "calc"
calc<-sampsize.oneway(tau=0.10,
                     delta=0.50,
                     theta=theta,
                     ngrps=2,
                     BH=TRUE,
                     n0=2,
                     nmax=50)

# Display the contents of "calc"
print(calc)
```