Simulated Correlation Experiment with Benjamini-Hochberg Procedure

Michael G. Nash

April 6, 2018

I wrote this code for a paid internship with the Biostatistics and Design Program at Oregon Health and Science University. I am its sole author. I am sharing it with the Github community with their permission.

The function 'sim.exp' simulates the results of experiments measuring associations between amounts of specific proteins in biological samples and a phenotype measurement, but could be used for prospective power analysis for any type of experiment in which one is testing for pairwise associations between one measure and many others. A permutation test is used to assign p-values to sample correlations based on a null hypothesis of no association between protein and phenotype. The Benjamini-Hochberg procedure is used to select significant associations between phenotype and protein measures at a designated false discovery rate (FDR).

The number of subjects, proteins, true positives, and times the experiment is repeated, the desired FDR, the strength of the association for those proteins associated with the phenotype, and the type of correlation being measured (Spearman or Pearson) can be set by arguments to the function 'sim.exp'. Simulated data are bivariate normal (i.e. normal distribution of protein and phenotype with linear association) with no missing values. For each simulated experiment, sim.exp returns the number of true and false positives, and it is easy to calculate from these the observed proportion of false positives.

In the following example, I simulate 100 experiments each with 100 subjects each in which 600 proteins are measured, 30 of which are associated with the phenotype in the population, under 2 FDRs and with 3 different levels of correlation. Summary statistics are shown for counts of true and false positives and the proportion of false positives.

```
## [1] "                                FDR = 0.1 , rho = 0.1"
## [1] "True Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.08    0.00    2.00
## [1] "False Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.08    0.00    2.00
## [1] "Proportion False Discoveries:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  1.0000  0.5385  1.0000  1.0000      87
## [1] "                                FDR = 0.1 , rho = 0.3"
## [1] "True Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   12.00   17.00   16.56   21.00   28.00
## [1] "False Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    3.00    3.13    4.00   11.00
## [1] "Proportion False Discoveries:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.08173 0.13961 0.15301 0.21147 0.66667
## [1] "                                FDR = 0.1 , rho = 0.5"
## [1] "True Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.00   30.00   30.00   29.74   30.00   30.00
## [1] "False Positives:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.0     4.0     4.2     6.0    13.0
## [1] "Proportion False Discoveries:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0625  0.1176  0.1190  0.1667  0.3023
## [1] "                                FDR = 0.2 , rho = 0.1"
## [1] "True Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.21    0.00    2.00
## [1] "False Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.54    1.00    9.00
## [1] "Proportion False Discoveries:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.5000  1.0000  0.6913  1.0000  1.0000      60
## [1] "                                FDR = 0.2 , rho = 0.3"
## [1] "True Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   14.00   17.00   16.67   20.00   26.00
## [1] "False Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00    4.00    4.37    6.00   11.00
## [1] "Proportion False Discoveries:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1250  0.1952  0.1959  0.2682  0.4211
## [1] "                                FDR = 0.2 , rho = 0.5"
## [1] "True Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   30.00   30.00   29.87   30.00   30.00
## [1] "False Positives:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    7.00    9.00    9.67   13.00   29.00
## [1] "Proportion False Discoveries:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09091 0.19313 0.23077 0.23774 0.30233 0.49153
```

To use the function 'sim.exp':

Run 'simulate.R'. The functions 'sim.exp' and 'getcdf' are now part of your global environment. 'getcdf' is used by 'sim.exp' to assign p values to sample correlations. 'sim.exp' takes the following arguments:

nrep: number of simulated experiments
nsubj: number of subjects in each experiment
nprot: number of protein measurements in each experiment
ntrue: number of protein measurements associated with the phenotype in the population (from zero to nprot)
FDR: the false discovery rate one wishes to use (from 0 to 1)
spearman: correlations are measured using Spearman correlation if TRUE, Pearson otherwise

The defaults are as follows: nrep=1,nsubj = 16,nprot=600,ntrue=30,FDR=.5,rho=.5,spearman=TRUE

'sim.exp' returns a data frame containing the number of true and false positives for each experiment, labeled as 'true_positive' and 'false_positive', respectively. To get basic summary statistics for results of simulated experiments without saving results, use the following syntax:

summary(sim.exp())

I encourage researchers to use my code freely, as long as they credit its author, Michael G. Nash.