Homework 1
BIOS 7659/CPBS 7659
Due 9/18 in class

Please use one of the reproducible research templates (e.g., R Markdown, knitr, sweave, etc) and return your homework in report format.

1. Experimental Design Case Study

   Read attached files "CaseStudy3.pdf" and "CaseStudy4.pdf". For each case study, compare and contrast the different experimental designs with respect to the issues we discussed in class (resources and cost, variability, confounding, *etc.*). For each case study, write your answers in the form of a report to give to a collaborator suggesting which experimental design(s) are appropriate. Relate the report back to the specific applications in the case studies.

2. Sample Size Calculations

   Your collaborator is writing a grant and has proposed the use of microarrays to investigate kidney gene expression differences between treated and non-treated mice. She has asked you to help her with the power analysis. In particular, she needs an estimate of the sample size and budget for performing her experiments.

   Read section 3.8 in MA (also on Canvas). For several different power levels (.80 to .95) determine the cost of performing her experiment if she would like to detect a two-fold difference between the treated and untreated mice with a significance level of .0001 and an array that costs $1000 and has 20,000 probe sets. Assume that the Affymetrix GeneChip platform will be used (single-channel). Use the appropriate estimates of variability given in section 3.8.

   Summarize your calculations in a table and write up your results in a report format for her to put in the grant submission. Include the assumptions, null and alternative hypothesis, statistical test and expected number of false positives. Include details on your calculations (software version, functions, code, *etc.*).

   Note: You can calculate the sample size using equation 3.2 in MA or you can use `pwr.t.test()` in the `R pwr` package from CRAN.

3. Sample Size Comparisons

   - You will compare different methods for determining the sample size. Each part should only be a few lines of code.
   - You will need several R packages for this problem
     - Install the package `pwr` from CRAN,
     - Install the package `ssize` from Bioconductor:

       ```
       source("http://bioconductor.org/biocLite.R")
       biocLite("ssize")
       ```

– The `samr` package is not being maintained, but you can download from this site, and put in your local directory.

  `https://cran.r-project.org/src/contrib/Archive/samr/samr_2.0.tar.gz`

  Then, follow these steps to install (but change your local directory)

```
source("http://bioconductor.org/biocLite.R")
biocLite("impute")
install.packages(c("matrixStats", "GSA", "shiny", "openxlsx"))
install.packages("Desktop/7659/homeworks/hw1/samr_2.0.tar",
      repos = NULL, type="source")
```

Three files will be provided (`sdvalues.txt, pvalues.txt, arraydata.txt`) and are calculated from a sample data set of human cell lines treated with low or high levels of a treatment (n=4 per group). The data are from the Affymetrix GeneChip platform and have already been log transformed and summarized at the probeset level.

(a) What is the sample size needed based on $\alpha = .001$, fold change of 2 ($\delta = 1$ in $\log_2$) and standard deviation of 0.5 to achieve power of at least 0.8 or 0.95? Use `pwr.t.test()` in the `pwr` package. Summarize your findings, and explain π_0.

Note: The d option in `pwr.t.test()` is δ/sd.

(b) As in part a), determine the sample size needed, but with a FDR of 0.05 instead. Use `power.t.test.FDR()` in the `ssize` package. Summarize your findings. Explain π_0.

(c) The attached file `sdvalues.txt` contains pooled standard deviations (for the two groups) for the example data set. Read this file and plot the density (or histogram) of the standard deviations. Use `ssize()` and `ssize.plot()` in the `ssize` package to examine the sample size based on these standard deviations. Use `sig.level`, `delta` and `power` as in part a). What do you conclude from the plots?

(d) The attached file `pvalues.txt` contains the p-values from a two-sample t-test for each probeset from the example data set. Use the PowerAtlas online tool to examine the sample size based on these p-values (be patient, the server is slow and it could take 5-10 minutes). What do you conclude from the plots?

`http://poweratlas.ssg.uab.edu/`

(e) The attached file `arraydata.txt` contains the data for each sample. Use `samr.assess.samplesize()` in the `samr` package, which implements the method from journal club (Tibshirani, 2006), to examine the sample size based on these data. What plots are displayed with `samr.assess.samplesize.plot()`? What do you conclude from the plots?

Hint: To set up a `samr` object in R:

```
>x = as.matrix(read.table("arraydata.txt", row.names = 1, header = TRUE))
```

```
>data = list(x=x,y=c(rep(1,4),rep(2,4)), geneid=row.names(x), genenames
= row.names(x), logged2 = TRUE)
>samr.obj = samr(data, resp.type="Two class unpaired", nperms=100)
```

(f) Describe the different methods and then compare and contrast the sample size calculations in parts a)-e).