Homework 1 Solutions
BIOS 7659/CPBS 7659

1) Experimental Design Case Study

a) Case Study 3 (from Stekel 2003)

Introduction: A microarray study will be performed to identify subtypes of patients suffering from B-cell lymphomas based on their gene expression profiles. Three experimental designs have been proposed for a study of 60 patients.

Methods: Design 1 uses a two-color array, where the 60 samples are hybridized in pairs (one Cy3, the other Cy5). Design 2 also uses a two-color array but each patient is hybridized (Cy3) with a universal reference sample (Cy 5). Design 3 uses the Affymetrix single color array and each sample is hybridized to its own array.

Results/Discussion: With respect to arrays and materials, Design 1 uses the least arrays (30) compared to the others (60). Designs 1 and 3 use the least samples compared to Design 2, which needs a universal reference for each array. Design 1 is not a good design because we cannot compare subjects on equal footing since they are always hybridized in pairs. It is difficult to make comparison between two samples hybridized on different arrays, particularly if they are labeled with different dyes. Although this design uses fewer arrays, this is not a good design for the purpose of subtyping, which requires direct comparison of gene expression profiles between patients. In contrast, Design 2 is better suited for direct comparisons between patients since they are always compared to a universal reference and they are all hybridized with the same dye. Design 3 is also suited for direct comparisons. Uniformity of the Affymetrix platforms makes comparisons between samples meaningful. Subject and array effects are confounded but the arrays can be normalized using between-array normalization so that they can be compared. Compared to Design 2, Design 3 also does not need a universal reference. Based on the advantages of Design 3, we recommend that design.

b) Case Study 4 (from Stekel 2003)

Introduction: A microarray experiment will be performed to identify genes in budding yeast that show similar expression profiles during sporulation, which is the process where budding yeast can reproduce sexually by producing haploid cells.

Methods: Three experimental designs have been proposed for the time course. Design 1 uses the Affymetrix array where a single sample from seven time points (time 0 -6) are hybridized to each array. Design 2 uses a two-color array where each time point (time 1-6) is hybridized (Cy3) with time point 0 as a reference (Cy5). Design 3 also uses a two-color array in a loop design, where each sample is labeled twice, one with Cy3 and once with Cy5.

Results/Discussion: With respect to arrays and materials, Design 2 uses the least arrays (6) compared to the others (7). Designs 3 use the most material since each time point is hybridized twice.

Design 1 has a serious problem if one of the arrays is ‘brighter’ than the others. This could be due to overall gene expression being higher, resulting from biological differences, or due to technical/array artefacts. These two factors are confounded and no normalization or analysis can identify the true situation. However, multiple biological samples per time point would resolve this problem. Alternatively Design 2 and 3 resolve this problem. In Design 2, each sample is normalized relative to time 0. So if the array is ‘brighter’, it will be bright for both samples hybridized, so the log ratio is free from this artefact. Design 3 similarly resolves this problem and also has two technical replications per time point. However, Design 3 has two disadvantages compared to Design 2. First, it is difficult to analyze because of the paired time points and will required a more complex analysis. Second, if one array fails, it will affect the entire analysis because of the pairing. In contrast for Design 2, a failed array can be omitted from the analysis so that the other data can be used. Based on the advantages of the different methods, we recommend Design 2, or using more samples in Design 1.

2) Sample Size Calculations

This study will investigate gene expression differences in the kidney between treated and non- treated mice. For each gene, the null hypothesis is that there are no expression differences between the two groups compared to the alternative hypothesis that there are expression differences between the two groups. To calculate the necessary power needed in the study, calculations were based on a proposed two-fold difference in log expression levels between classes, a significance level of 0.0001, and a standard deviation of 0.25 units. The standard deviation is the approximate median value observed for inbred mice gene expression using the Affymetrix GeneChip [1]. The necessary sample size needed to achieve the desired power in the study is listed below for a two-sided two-sample t-test with equal variance between the groups.

The power represents the probability that the alternative hypothesis will be chosen given it is indeed true. These calculations assume the expression measurements are independent across genes and are approximately normally distributed among samples of the same class. The expected number of false positives in this model is low at 2 (.0001 X 20,000), which translates to the expected number of genes declared significant when they are actually not differentially expressed. The expected costs of arrays for performing the analysis are also listed below.

Sample Size (per group)	*Power*	*Cost*
7	0.80	14000
7	0.85	14000
8	0.90	16000
8	0.95	16000

For $14,000 a sample size of 7 per group (14 microarrays) will give adequate power (85%). If the budget permits an additional $2000, we suggest a sample size of 8 per group, which will provide up to 95% power to detect at least 2-fold differences in gene expression between the two groups.

```
Code:
require(pwr)
alpha <- 0.0001; sd <- 0.25
delta <- 1; d <- delta/sd
power <- 0.75
n <- NULL

for (i in 1:4) {
     power <- power + 0.05
     tmp <- pwr.t.test(d=d, sig.level=alpha, power=power)

     tmp <- ceiling(tmp$n)
     n <- rbind(n, c(tmp, power, tmp*2*1000))
}
n
```

[1] Simon R, Korn E, McShane L, Radmacher M, Wright G, Zhao Y. *Design and Analysis of DNA Microarray Investigations*, Springer-Verlag, New York, 2003.

3) Sample Size Comparisons

We will compare five methods for sample size calculations for the proposed study in problem 2. This study will investigate gene expression differences in the kidney between treated and non- treated mice. For each gene, the null hypothesis is that there are no expression differences between the two groups compared to the alternative hypothesis that there are expression differences between the two groups.

a) A sample size of 12 per treatment group (24 in total) will achieve at least 80% power and a sample size of 15 per treatment group (30 in total) will achieve at least 95% power to detect a two-fold change between treated and untreated groups with an estimated standard deviation of 0.5 and a significance level of 0.001 using a two-sided two-sample t-test assuming equal variance.

Code:

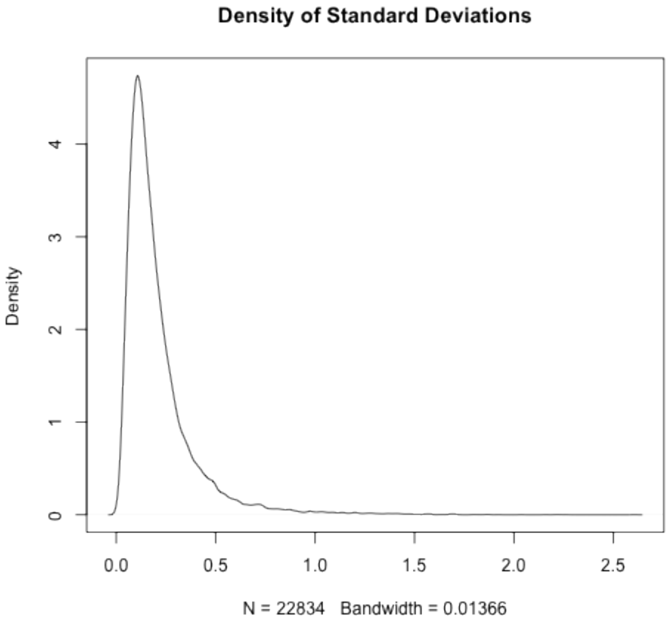
```
library(pwr)

ceiling(pwr.t.test(d=1/.5,power = .8, sig.level=0.001, type="two.sample",
alternative="two.sided")$n)

ceiling(pwr.t.test(d=1/.5,power = .95, sig.level=0.001, type="two.sample",
alternative="two.sided")$n)
```

b) If we assume a prior (i.e., before observing the data) probability of 0.95 for the fraction of genes that are not differentially expressed (π_0), then a sample size of 11 per treatment group (22 total) will achieve at least 80% power, and a sample size of 14 per treatment group (28 total) will achieve at least 95% power to detect a two-fold change between treated and untreated groups. These calculations are based on an estimated standard deviation of 0.5 and a false discovery rate (FDR) of 0.05 using a two-sided two sample t-test.

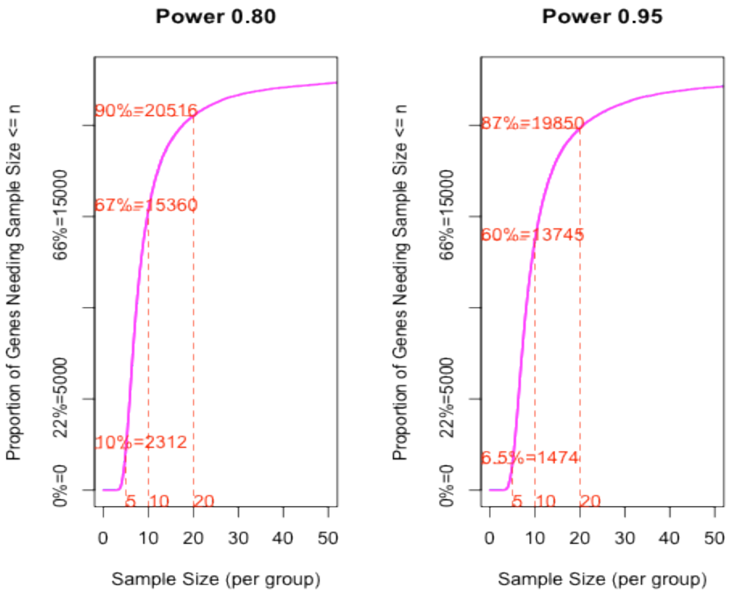
Code:

```
library(ssize)

ceiling(power.t.test.FDR(sd=.5, delta=1,  FDR.level=0.05, pi0=0.95, power=.80,
type="two.sample", alternative="two.sided")$n)

ceiling(power.t.test.FDR(sd=.5, delta=1,  FDR.level=0.05, pi0=0.95, power=.95,
type="two.sample", alternative="two.sided")$n)
```

c) The density of standard deviations from a pilot study is displayed below. The mode of the distribution is below 0.25 (which is lower than the value used for part b) and the plot shows a right skewed distribution.



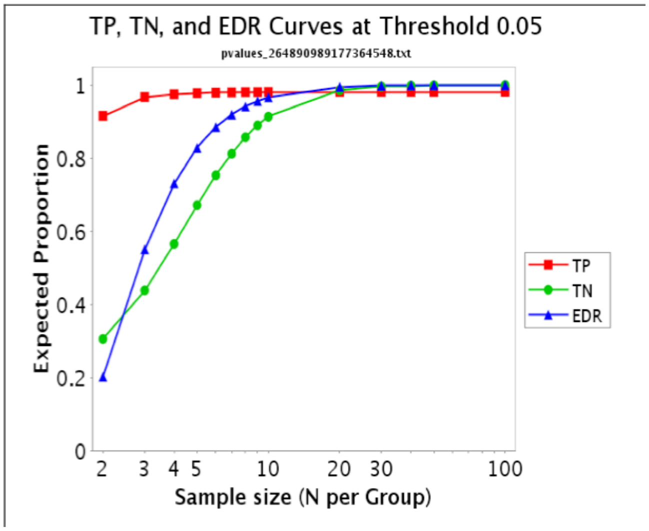
Based on the `ssize` package, the resulting plots show the relationship between sample size and the number of genes that reach the designated power. A sample size of 20 per group is required to ensure that at least 90% of genes have power of at least 80% (left plot) to detect a two-fold change between treated and untreated groups with the observed standard deviations and a significance level of 0.001. A sample size of 10 per group is sufficient if it is only required that 67% of the genes need to achieve 80% power. For 95% power (right plot), a sample size of 20 per group is required to ensure that at least 87% of the genes have power greater than 95% to detect a two-fold change between treated and untreated groups with the observed standard deviation and a significance level of 0.001. A sample size of 10 per group is sufficient if it is only required that 60% of the genes need to achieve 95% power.



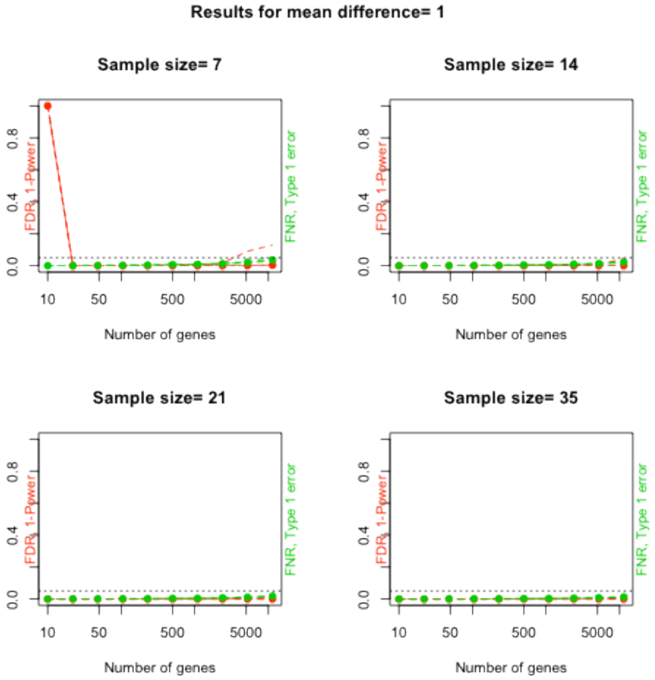
Code:

```
fold.change=2.0; sig.level=0.001;
sds = read.table("sdvalues.txt", row.names = 1)

plot(density(unlist(sds)), main = "Density of Standard Deviations")

pwr=0.8;
all.size.80 <- ssize(sd=sds[,1], delta=log2(fold.change), sig.level=sig.level,
power=pwr)
ssize.plot(all.size.80, lwd=2, col="magenta", marks = c(2,5,10,20,50), xlim =
c(0,50), main = paste("Power ", pwr, collapse = ""))

pwr=0.95;
all.size.95 <- ssize(sd=sds[,1], delta=log2(fold.change), sig.level=sig.level,
power=pwr)
ssize.plot(all.size.95, lwd=2, col="magenta", marks = c(2,5,10,20,50), xlim =
c(0,50), main = paste("Power ", pwr, collapse = ""))
```

d) In the PowerAtlas analysis, the expected discovery rate (EDR) is the proportion of genes that are truly differentially expressed (non-null) that will be called significant at the chosen significance level. This is analogous to average power. The probability of a true positive (PTP) is the expected proportion of genes that are called significant that are truly differentially expressed between the two groups, which is related to the FDR. The probability of a true negative (PTN) is the expected proportion of genes that are not called significant that are truly not differentially expressed between the two groups. Higher values for all three rates are more desirable. The results from our data set indicate that at an alpha level of 0.05 and a sample size of 10 results in all rates above ~0.90. At a sample size of 20, the rates are all very close to 1.0.



e) The plot from the SAM analysis displays the median permutation false discovery rate (FDR) in red and false negative rate (FNR) in green. The dashed lines are the 10th and 90th percentiles of the permutations. The plot below indicates that for sample size of at least 14 per group the FDR and FNR both fall below the horizontal black dotted line drawn at 0.05 for detecting a 2-fold change.



```
Code:
library(samr)
x = read.table(“arraydata.txt", row.names = 1, header = TRUE)
data=list(x=x,y=c(rep(1,4),rep(2,3)), geneid=row.names(x), genenames =
       row.names(x), logged2 = TRUE)
samr.obj<-samr(data, resp.type="Two class unpaired", nperms=100)
samr.assess.samplesize.obj<- samr.assess.samplesize(samr.obj, data, log2(2))
samr.assess.samplesize.plot(samr.assess.samplesize.obj)
```

f) The methods in part (a)-(c) offer more traditional power calculations for two sample t-tests but two of the approaches are better suited for microarray data by either using an FDR correction (b) or using a distribution of standard deviations based on pilot data (c). Method (d) relies on parametric bootstrapping on pilot data of two-sample test p-values, while method (e) relies on permutations and does not require independence among genes. If pilot data is available, then methods (c)-(e) are applicable otherwise (a)-(b) offer more standard alternatives. Despite the differences among the methods, they all indicate that a sample size of 10-15 will provide adequate power for the proposed study.