Stat 516 Homework 6 9:30 A.M., Tuesday, April 26

1. Visit the Gene Expression Omnibus site at http://www.ncbi.nlm.nih.gov/geo/ and find an experiment that looks interesting to you. One way to do this is to click the "Series" link under GEO navigation, Browse, and GEO Accessions on the main page. This will take you to a list of several thousand GSE accessions. Clicking a GSE number will take you to a page with information about the dataset and links to the data. Once you have selected a dataset, complete the following parts. Provide the code you used to get your answers along with your answers.

   (a) Give the GSE number of the dataset that you selected.

   (b) Provide a brief description of the experiment. You may cut and paste from the information provided on GEO.

   (c) Describe one test of interest that you will perform for each gene (e.g., for example, test for differential expression between treatment 1 and treatment 2). Normalize the data and then use limma to perform the test for each gene.

   (d) Give estimates of the prior degrees of freedom and the prior variance produced by limma.

   (e) Make a figure analogous to slide 35 of slide set 19 to show how well the estimated marginal distribution of the original variance estimates matches the empirical distribution of the original variance estimates.

   (f) Give the original variance estimate and Smyth's empirical Bayes estimate of the variance for gene 1.

   (g) Present a histogram of p-values (from the test in part c) with a plot of the estimated uniform-beta mixture fit, and give the estimated parameters for the best uniform-beta fit.

   (h) Convert the p-values into q-values and provide the number of significant genes for FDR control at levels 0.01, 0.05, and 0.10.

   (i) List the gene IDs of the 10 genes with the smallest p-values.

   (j) Find estimates of the posterior probability of differential expression for each gene. Give the p-values along with the estimated posterior probabilities of differential expression for the 5 genes with the smallest p-values and the 5 genes with the largest p-values.

2. Suppose you will test 20,000 six-sided dice in search of dice that have a probability of rolling 6 that is greater than $1/6$. Your plan is to roll each die four times and declare any die that rolls 6 all four times to be a die that has probability of rolling 6 that is greater than $1/6$. Suppose that, unknown to you, one die will roll 6 with probability 1, 10 dice will roll 6 with probability 0.5, 20 dice will roll 6 with probability 0.4, and 100 dice will roll 6 with probability 0.2. The other 20,000-(1+10+20+100) dice are regular six-sided dice that roll 6 with probability $1/6$. Use the definitions given in the notes on mixture modeling of the p-value distribution to compute the following quantities for this die-rolling scenario. Of course, you will need to draw an analogy between this hypothetical die testing problem and testing for differential expression in order for this problem to make sense (e.g., regular dice are like equivalently expressed genes, die with greater than $1/6$ probability of landing heads are like differentially expressed genes, etc.).

   (a) FDR

   (b) TPR

   (c) TNR

(d) EDR

3. Complete all parts of this problem without the use of a computer to make sure that you understand the details of the clustering algorithms. Consider the following "data" to be clustered as described below.

$$10 \quad 20 \quad 40 \quad 80 \quad 85 \quad 121 \quad 160 \quad 168 \quad 195$$

For each part of the problem, assume that Euclidean distance will be used to measure the distance between the data points.

(a) Use hierarchical agglomerative clustering with single linkage to cluster the data. Draw a dendrogram to illustrate your clustering and include a vertical axis with numerical labels indicating the height of each parental node in the dendrogram.

(b) Repeat part (a) using hierarchical agglomerative clustering with complete linkage.

(c) If two clusters are desired, what data points would be clustered together according to the single linkage method used in part (a)?

(d) If two clusters are desired, what data points would be clustered together according to the complete linkage method used in part (b)?

(e) Use the K-means algorithm with K=3 to cluster the data set. Suppose that the points 160, 168, and 195 were selected as the initial cluster means. Work from these initial values to determine the final clustering for the data. Show your work so that it will be easy to see each step you took to get from the initial values to your final clustering.

(f) Show that the clustering produced by the K-means algorithm depends on starting mean values by providing a different set of three starting means that results in a different set of final clusters. You need to provide only the starting means and the final clustering. It is not necessary to show your work.

4. Suppose the SAM method is used to identify significantly differentially expressed genes in a completely randomized two-treatment experiment where one Affymetrix GeneChip is used to measure expression in each experimental unit. Based on the selected Δ value, 39 genes exceeded the thresholds for significance.

(a) Use the method proposed by Tusher et al. (2001) to estimate the FDR associated with this list of 39 genes using the number of genes exceeding the thresholds in all possible permutations of the data provided below.

Permutation	Number of Genes Exceeding Thresholds
1	39
2	4
3	18
4	0
5	6
6	8
7	2
8	17
9	9
10	38

(b) If only 10 permutations of the data were possible, what were the samples sizes for the treatment groups?

5. Consider the parametric empirical Bayes method proposed by Kendziorski et al. (2003). For the two-treatment case, each gene can be classified into two groups (EE or DE). In the three-treatment case, each gene can be classified into 5 groups. How many groups for the four-treatment case?

6. Consider the work of Smyth (2004) on estimation of gene specific variance. Suppose a two-treatment completely randomized design has been conducted with 4 experimental units per treatment and one Affymetrix GeneChip per experimental unit. Suppose $d_0 = 5$ and $s_0^2 = 2$.

   (a) Find the expected value of σ_j^2 given that $s_j^2 = 1.2$.

   (b) Give Smyth's estimator of σ_j^2 given that $s_j^2 = 1.2$.

   (c) Suppose that the original two-sample t-statistic for the gene in part (a) and (b) was 3.75. Find the value of the moderated t-statistic.

   (d) Compute a p-value for the gene.

7. Simulate data from a two-treatment microarray experiment using the R code below. Then complete the following parts.

```
set.seed(91184)
z=matrix(rnorm(80000),nrow=10000)
sds=sample(c(runif(7000,.75,1),
             runif(2000,1,1.5),
             runif(1000,1.5,3)))
z=z*sds
trt.effects=rbinom(10000,1,.3)*(2*rbeta(10000,3,3)+.5)*sds
d=cbind(z[,1:4]+trt.effects,z[,5:8])
```

   (a) State the number of genes simulated (m), the sample size per treatment group (n), and the number of true null hypotheses (m_0).

   (b) What is the distribution of true gene-specific variances? Does this distribution match the prior assumed by the `limma` package?

   (c) Conduct a two-sample t-test for each gene, assuming equal variance across treatment groups within gene but different variances across genes. Give the p-values for genes 1 through 5.

   (d) Use `limma` to obtain a p-value for each gene and give the p-values for genes 1 through 5 obtained by the empirical Bayes approach.

   (e) A Receiver Operator Characteristic (ROC) curve plots sensitivity vs. 1 - specificity. Sensitivity is defined as the proportion of false null hypotheses that are rejected. Specificity is the proportion of true null hypotheses that are not rejected. Ideally a method would have both high sensitivity and high specificity. Given a list of p-values and information about whether the null or alternative hypothesis is true for each p-value, it is possible to create an ROC curve by varying the threshold for significance. As the threshold for significance changes, different (1-specificity, sensitivity) pairs are generated. Points corresponding to thresholds set at each of the observed p-values can be plotted and connected with line segments to form an ROC curve. Generate an ROC curve for the part (c) results and a second curve for the part (d) results. Plot both ROC curves on the same graph. (Use p-values for all genes, not just the five p-values reported in each of parts (c) and (d).)

   (f) Based on the ROC curves, which analysis method seems preferable for this simulated dataset?