Stat 416 Homework 6 Due 5:00 P.M., Friday, May 1

1. Suppose a test for differential expression is conducted for each of 100 genes. The following table provides information about the observed p-values.

Range	Number of p-values
[0.0-0.1]	22
(0.1-0.2]	10
(0.2-0.3]	18
(0.3-0.4]	10
(0.4-0.5]	9
(0.5-0.6]	6
(0.6-0.7]	4
(0.7-0.8]	7
(0.8-0.9]	8
(0.9-1.0]	6

   (a) Estimate the number of true null hypotheses using the histogram-based estimator described in course notes. (Don't use the iterative version of the algorithm. Use the simple equivalent approach based on finding the leftmost bin with a count less than or equal to its tail average.)

   (b) Estimate the number of true null hypothesis using the λ-threshold method discussed in course notes with $\lambda = 0.8$. You may assume that no p-value equals exactly 0.8 when computing the estimator.

   (c) Using the estimate in part (b) for the number of true null hypotheses, fill in the appropriate blank below. (Put a number on only one of the blank lines, and be as precise as possible with your answer.)

   The q-value for a gene with a p-value of 0.2 is

   - less than or equal to _____.
   - equal to _____.
   - greater than or equal to _____.

2. Suppose the following are log-scale expression measurements for a single gene from a two-treatment completely randomized experimental design.

Treatment 1		Treatment 2	
4.3	5.1	9.7	6.2

   Is there evidence that expression of the gene depends on the treatments? Conduct a permutation test to address this question by completing the following parts.

   (a) There are four observations total 4.3, 5.1, 6.2, and 9.7. Write down all the possible ways of assigning two of the four observation to treatment 1 and two to treatment 2. (Note that the order of the observations within treatment groups doesn't matter. Thus, 4.3, 5.1 vs. 6.2, 9.7 is the same as 5.1, 4.3 vs. 6.2, 9.7, for example.)

   (b) For each permutation of the data in part (a), compute the sum of the observations assigned to treatment 1.

   (c) Determine the number of permutations that yielded a sum in part (b) that was less than or equal to the treatment 1 sum obtained from the actual data.

(d) Take the number in part (c) divided by the total number of permutations in (a) to get a one-sided p-value. Double that number to get a two-sided p-value.

(e) Would your answer change if you had use a different test statistic like the difference between treatment averages or the usual two-sample t-statistic in place of the sum of treatment 1 observations?

(f) In this case, was it necessary to write out all the permutations and to compute a test statistic for each in order to determine the permutation p-value? Explain.

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. (Note that K-means is just like K-medoids except that a mean of the points in each cluster is computed at each iteration and used instead of a medoid.) 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.

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. 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

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

   (b) 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.

   (c) Compute a p-value for the gene using R.

6. Download any data set of your choice from the Gene Expression Omnibus Website. Describe the experimental design used to generate the data as well as you can given available information. (You may need to take a look at supplemental materials or the publication that utilized the data to determine this information.) Provide R code that can be used to

   (a) read the data,

   (b) perform any one hypothesis test of potential scientific interest for each gene,

   (c) make a histogram of the p-values,

   (d) estimate the number of true null hypotheses,

   (e) determine the number of genes declared to be differentially expressed for FDR thresholds of 0.05, 0.10, and 0.15.

   In addition to the URL, description of the experimental design, and code, provide the histogram for part (c) and numerical answers for parts (d) and (e).