STAT 516 HW6 Solutions

1. Answers vary.

2. a)

$$\begin{aligned}\text{FDR} &= P(EE|4\ \ 6s) = \frac{P(4\ \ 6s|EE)P(EE)}{P(4\ \ 6s)} \\ &= \frac{(1/6)^4(19869/20000)}{\{(1/6)^4 * 19869 + (1/5)^4 * 100 + (2/5)^4 * 20 + (1/2)^4 * 10 + 1^4 * 1\}/20000} \\ &= 0.869696\end{aligned}$$

b)

$$\text{TPR} = 1 - \text{FDR} = 0.130304$$

c)

$$\begin{aligned}\text{TNR} &= P(EE|\text{not } 4\ \ 6s) = \frac{P(\text{not } 4\ \ 6s|EE)P(EE)}{P(\text{not } 4\ \ 6s)} \\ &= \frac{(1-1/6^4)(19869/20000)}{[(1-1/6^4) * 19869 + (1-1/5^4) * 100 + \{1-(2/5)^4\} * 20 + (1-1/2^4) * 10]/20000} \\ &= 0.9935592\end{aligned}$$

d)

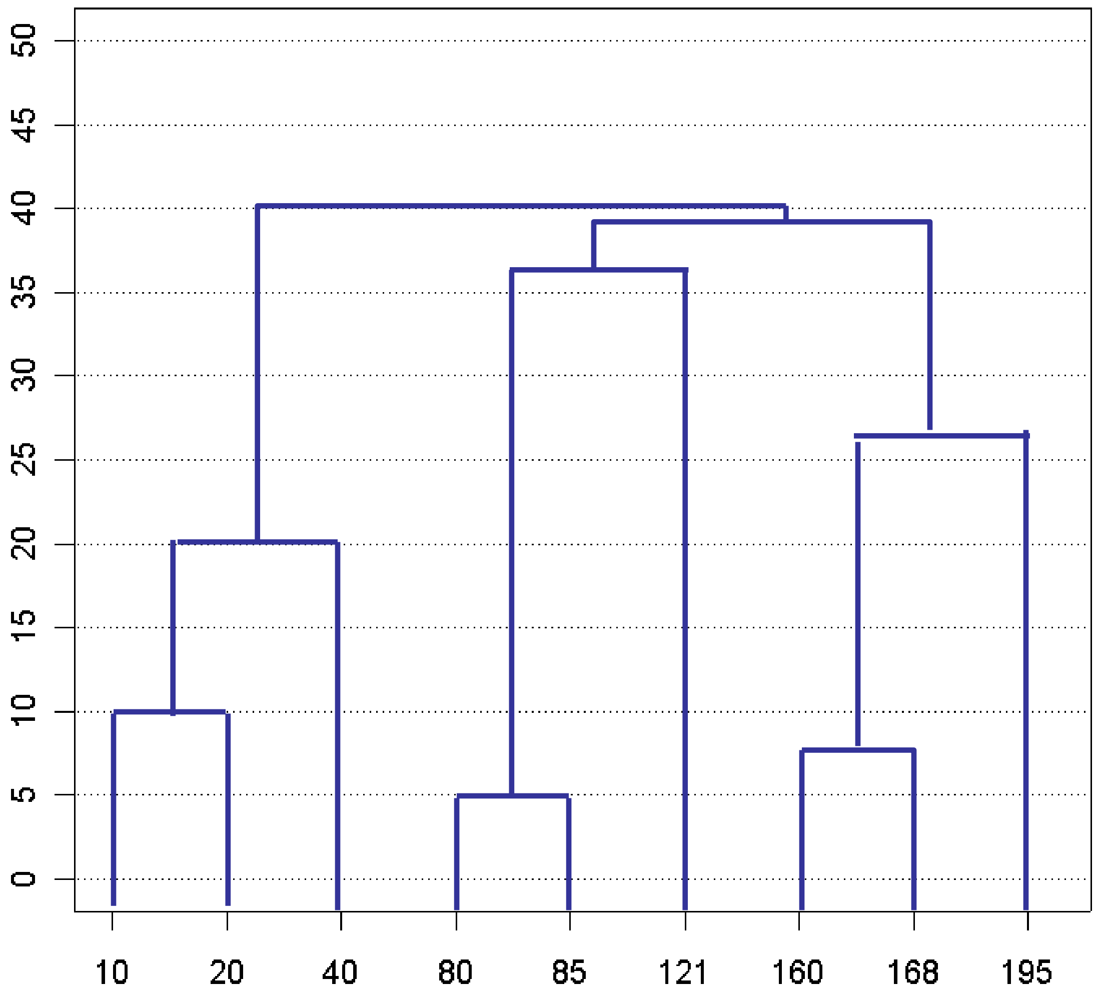
$$\begin{aligned}\text{EDR} &= P(4\ \ 6s|DE) \\ &= \frac{\{(1/5)^4 * 100 + (2/5)^4 * 20 + (1/2)^4 * 10 + 1^4 * 1\}/20000}{131/20000} \\ &= \frac{(1/5)^4 * 100 + (2/5)^4 * 20 + (1/2)^4 * 10 + 1^4 * 1}{131} \\ &= 0.01753435\end{aligned}$$

3. Consider the following "data" to be clustered using a variety of methods described below.

10 20 40 80 85 121 160 168 195

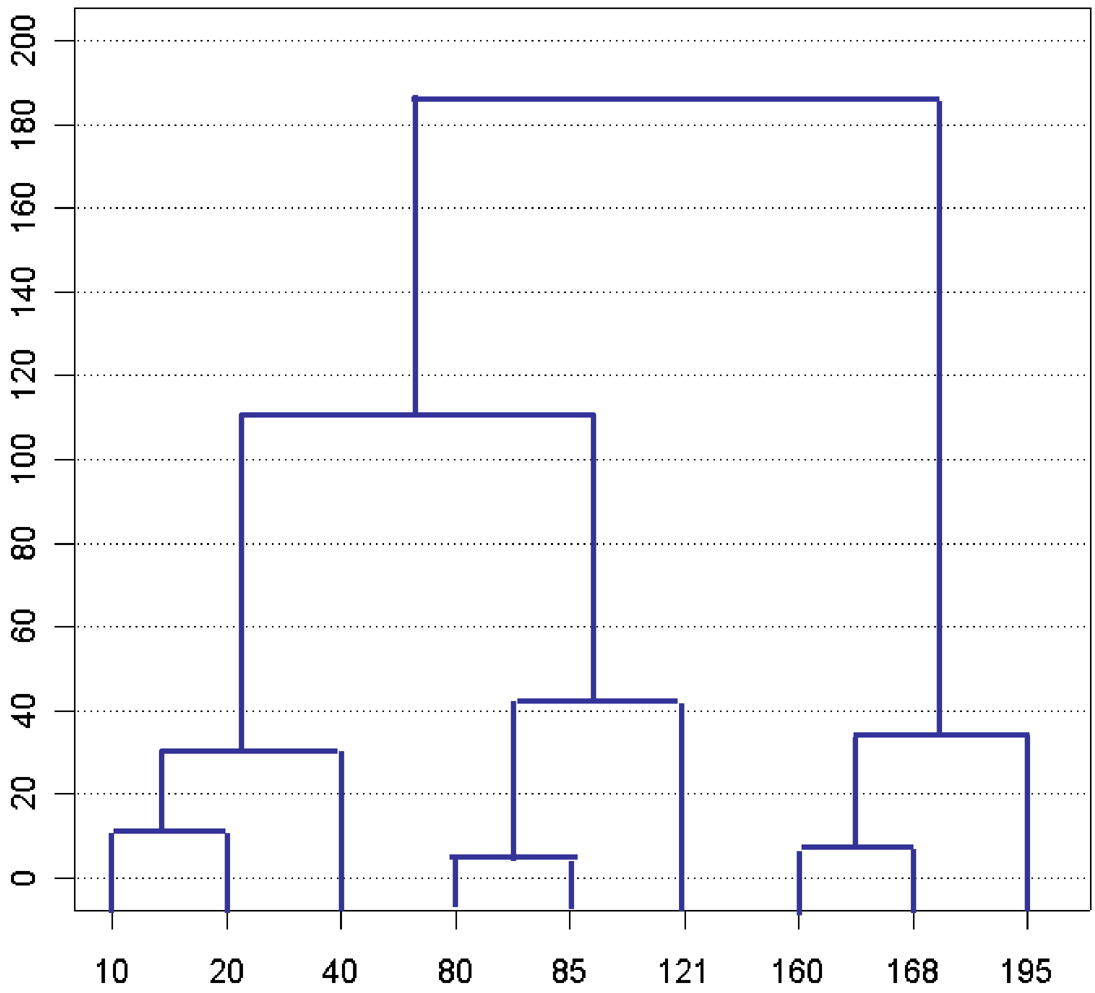
For each part of the problem, assume that Euclidean distance will be used to measure the distance between the data points.

a) Use hierarchical agglomerative clustering with single linkage to cluster the data. Draw a dendrogram to illustrate your clustering using the axes below.



3. (continued)

b) Use hierarchical agglomerative clustering with complete linkage to cluster the data. Draw a dendrogram to illustrate your clustering using the axes below.



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

(10, 20, 40) and (80,...,195)

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

(10,...,121) and (160, 168, 195)

3. (continued)

e) Use the K-means algorithm with K=3 to cluster the same data set that is copied below for your convenience.

10 20 40 80 85 121 160 168 195

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.

1. (10, 20, 40, 80, 85, 121, 160) (168) (195)

means: 73.71 168 195

2. (10, 20, 40, 80, 85) (121, 160, 168) (195)

means: 47 149.67 195

3. same as step 2.

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.

means: 23.3333 95.3333 174.3333

final clustering: (10, 20, 40) (80, 85, 121) (160, 168, 195)

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

mean of 39, 4, 18, 0, 6, 8, 2, 17, 9, 38 divided by 39 is 14.1/39=0.3615385

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

There are 10 ways to divide 5 objects into a group of 2 and 3 (5 choose 2 = 10). Thus the sample sizes must have been 2 for one treatment group and 3 for another.

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?

1 group where all treatments means the same (1234)
4 groups of the form (1)(234)
3 groups of the form (12)(34)
6 groups of the form (1)(2)(34)
1 group where all treatments are distinct (1)(2)(3)(4)

total of 15 groups

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.

d=4+4-2=6

(6*1.2+5*2)/(6+5-2)=1.9111

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

(6*1.2+5*2)/(6+5)=1.563636

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.

sqrt(1.2/1.563636)*3.75=3.285141

d) Compute a p-value for the gene.

Compare the value to a t-distributions with 6+5=11 d.f.

```
2*(1-pt(3.285141,11))
[1] 0.00726714
```

7.

a) m=10000, n=4

```
sum(trt.effects==0)
[1] 7003
```

Thus, m0=7003.

b) The true standard deviations are a mixture of uniforms. This is not the distribution assumed by limma (scaled inverse chi-square on variances).

c)

```
getp=function(y)
 {
  t.test(y[1:4],y[5:8],var.equal=T)$p.value
 }

 p2=apply(d,1,getp)

p2[1:5]
[1] 0.87182604 0.76423532 0.11908859 0.06168796 0.94096471
```

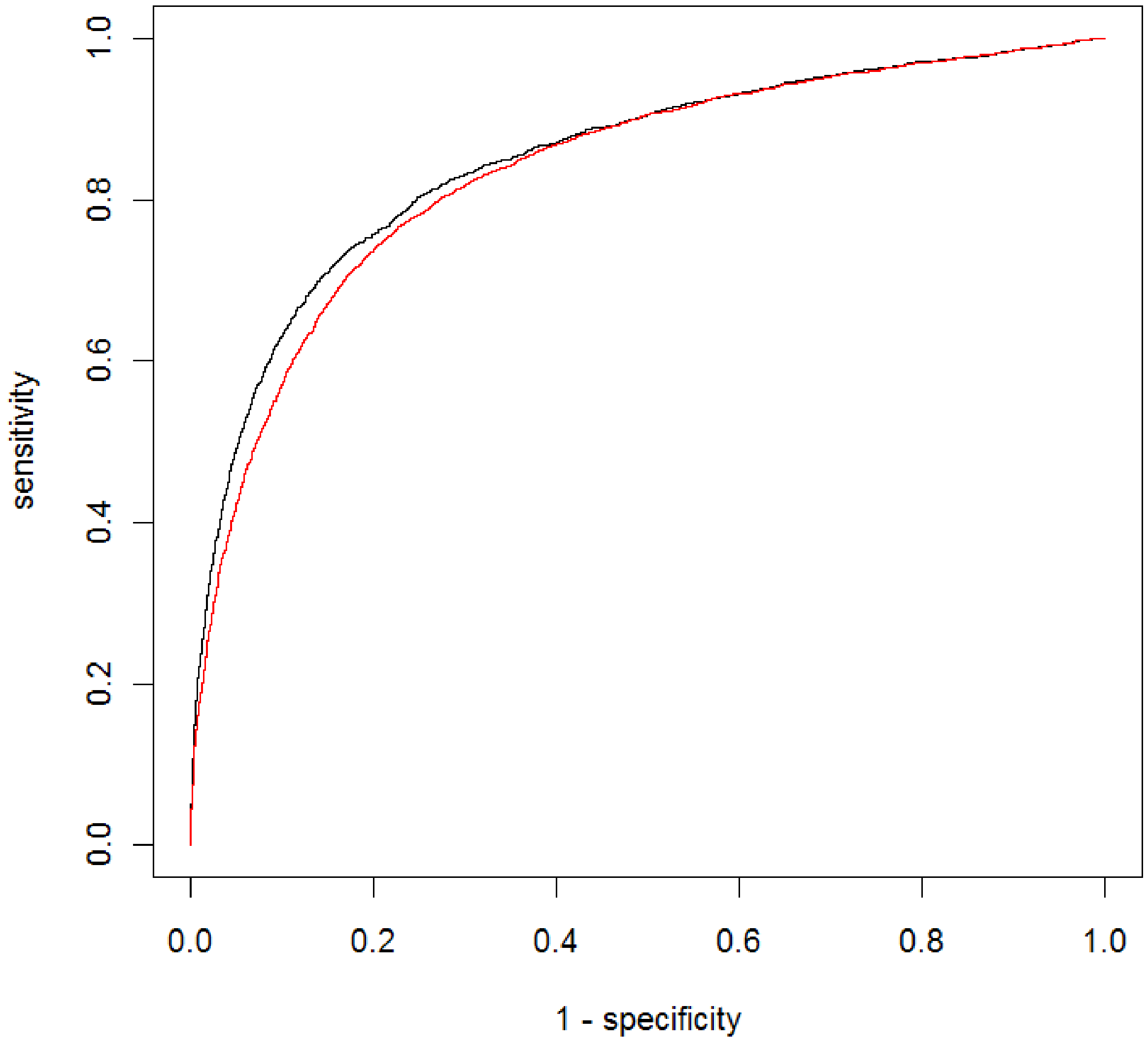
d)

```
library(limma)

trt=gl(2,4)
trt
design=model.matrix(~trt)
design
colnames(design)=c("mu","trtdiff")

fit=lmFit(d,design)
contr.mat=makeContrasts(trtdiff,levels=design)
fit2=contrasts.fit(fit,contr.mat)
fit3=eBayes(fit2)

fit3$df.prior
fit3$s2.prior
```

```
p=fit3$p.value[,1]
p[1:5]
[1] 0.85023917 0.83759893 0.12354957 0.06081275 0.92738212
```

e)

```
roc=function (x,truth,add=F)
{
  y=truth[order(x)]
  ones=sum(y)
  zeros=sum(1-y)
  sensitivity=cumsum(y)/ones
  specificity=1-cumsum(1-y)/zeros
  if(add){
    lines(1-specificity,sensitivity,col=2)
  }
  else{
    plot(1-specificity,sensitivity,type="l")
  }
}

de=trt.effects>0
roc(p,de)
roc(p2,de,add=T)
```



f) The empirical Bayes method was superior despite the fact that the prior distribution was not correct.