Stat 416 Homework 5

Due Date: Tuesday, April 21

Using the data set from Homework 4, answer the following questions.

1. Given the available data, it makes sense to treat this experiment as a completely randomized experiment with two factors: diet and batch. The researchers reported that very few genes exhibited a significant interaction between diet and batch.

   (a) Explain in words what an interaction between diet and batch would mean.

   (b) Test each gene for interaction between diet and batch. The code

   ```
   lm(y~a+b+a:b)
   ```

   will fit a linear model with factor a and b main effects as well as $a \times b$ interaction. The p-value for the test of $a \times b$ interaction can be found in the ANOVA table. Sketch the shape of the resulting p-value distribution or provide the actual histogram.

   (c) Do you agree with the researchers claim that few genes exhibited interaction between diet and batch? Explain.

2. Regardless of your answer to the previous question, from now on, consider only the additive model

$$y_{ijk} = \mu + \delta_i + \beta_j + \varepsilon_{ijk} \quad (i = 1, 2, 3, 4; j = 1, 2; k = 1, 2, 3)$$

   where μ denotes the model intercept; $\delta_1, \ldots, \delta_4$ denote diet effects; β_1 and β_2 denote batch effects; the ε_{ijk} values are independent and identically distributed $N(0, \sigma^2)$ random variables; and the y_{ijk} values denote the normalized log expression values for a given gene. Write R code to fit this model separately for each gene using the `lm` function. As part of each analysis, obtain the following:

   (a) an estimate of σ^2 (mean square for residuals),

   (b) a p-value for testing diet main effects,

   (c) a p-value for tesing batch main effects,

   (d) a p-value for all possible pairwise comparisons between diets. The pairwise comparisons between diets should be done by averaging the diet means over batches. For example, the null hypothesis for testing diet 3 vs. diet 4 should be

$$H_0 : (\mu + \delta_3 + \beta_1 + \mu + \delta_3 + \beta_2)/2 = (\mu + \delta_4 + \beta_1 + \mu + \delta_4 + \beta_2)/2,$$

   which can be simplified considerably with a little algebra.

3. Using the functions available at

   `http://www.public.iastate.edu/~dnett/microarray/multtest.txt`

to complete the following.

(a) For each of the eight tests conducted in part 2, estimate the number of true null hypotheses among all the genes tested.

(b) Convert each of the eight sets of p-values into q-values using the function `jabes.q`. For the test of the McDonald's diet vs. the mouse diet, report the number of gene declared to be differentially expressed at FDR levels of 0.01, 0.02, 0.03, 0.04, 0.05, and 0.10.

(c) Fit a mixture of a uniform distribution and a beta distribution to the p-values from the comparison of the chimp diet with the McDonald's diet. Report an estimate of the proportion of null genes and the estimated parameters of the beta distribution from this fit.

(d) Estimate the posterior probability of differential expression for each gene for the comparison of the chimp diet with the McDonald's diet. Report the number of genes with posterior probabilities of differential expression above 75% for this test.

4. Use the `limma` function to fit the additive linear model described in part 2. Obtain an empirical Bayes estimate of the variance σ^2 for each gene and a p-value for the comparison of the McDonald's diet with the mouse diet. Report the following:

(a) Estimates of d_0 and S_0^2,

(b) the proportion of genes for which the empirical Bayes estimate of the variance is less than the usual linear model estimate obtained in part 2 (a),

(c) the degrees of freedom used for the modified t-test comparing the McDonald's diet with the mouse diet,

(d) for the test of the McDonald's diet vs. the mouse diet, the number of gene declared to be differentially expressed at FDR levels of 0.01, 0.02, 0.03, 0.04, 0.05, and 0.10, when the p-values from the modified t-test are converted to q-values using the `jabes.q` function.