High Dimensional Data analysis (BIOS:7240)
Breheny

Assignment 2
Due: Tuesday, February 11

1. *Relationship between FDR and local fdr.* Derive the relationship between FDR and local FDR on slide 17 of the "Local false discovery rates" notes:

$$\mathbb{E}\{\text{fdr}(z)|z \in \mathcal{Z}\} = \text{Fdr}(\mathcal{Z}),$$

where $\text{fdr}(z)$ is the local FDR at point z and $\text{Fdr}(\mathcal{Z})$ is the FDR over the set $\mathcal{Z}$. To be clear, this problem involves the true FDR and local FDR – no estimates are involved.

2. *Scaled χ^2.* Show that if a random variable X satisfies $cX \sim \chi^2_\nu$, then

$$X \sim \text{Gamma}\Big(\frac{\nu}{2}, \frac{c}{2}\Big),$$

where $\text{Gamma}(\alpha, \beta)$ denotes the gamma distribution with shape parameter α and rate parameter β.

3. *FDR accuracy in the presence of correlated tests.* Simulate 1,000 z-statistics according to the following scheme:

$$\begin{aligned} Z_i &\sim \text{N}(3,1) && \text{for } i = 1, 2, \ldots, 100 \\ Z_i &\sim \text{N}(0,1) && \text{for } i = 101, 102, \ldots, 1000 \\ \text{Cor}(Z_i, Z_j) &= 0.1 && \text{if } i, j > 100 \\ \text{Cor}(Z_i, Z_j) &= 0 && \text{otherwise} \end{aligned}$$

Use the Benjamini-Hochberg procedure with the **known value** $\pi_0 = 0.9$ (see part **??**) to reject as many hypotheses as possible subject to an FDR cutoff of 20%, then calculate the actual false discovery rate among those rejected hypotheses.

   (a) Using the code `p.adjust(p, method='BH')` to calculate the Benjamini-Hochberg q values will not utilize the known value of π_0. Can we modify the output of the function in order to reflect the known value? If so, how? Or must we re-implement the entire Benjamini-Hochberg procedure ourselves?

   (b) Repeat the procedure described above 2,000 times. Calculate the average true FDR and the standard deviation of the true FDRs.

   (c) Same as (a), but without any correlation between the z-statistics (i.e., $\text{Cor}(Z_i, Z_j) = 0 \,\forall\, i, j$). Again, what are the mean and standard deviation of the true FDRs across the 2,000 replications?

   (d) Produce a figure (e.g., a pair of histograms or a boxplot) comparing the true FDRs from (a) and (b), and comment on the effect of correlation upon FDR control.

4. *Combining left and right tails.* In the *V. parahaemolyticus* surface-sensing analysis that we carried out in the "Hierarchical models" lecture, there was a considerable difference in the number of significant genes in the right and left tails of the distribution.

(a) Applying **fdrtool** to the results of the moderated t-tests (using **limma**) we obtained in that lecture, estimate the local false discovery rate of each gene in two ways. First, supplying all the p-values in a single analysis (note that this treats the left and right tails the same). Second, supplying p-values from the left and right tails separately (NOTE: do not supply z-values, or **fdrtool** will estimate the null).

For your answer, provide a 2×3 table counting the number of genes declared significant at threshold of fdr < 0.15, where the rows are symmetric/asymmetric and the columns are left/right/total.

(b) In the asymmetric analysis, a p-value might be considered significant if it is in the right tail, but not the left. How would you explain/justify this to a collaborator who wants to know why gene A is significant but gene B isn't, even though they have the same p-value?

(c) Comment on which analysis you consider to be more appropriate with respect to both power and false positives in this situation.

5. *Prostate cancer study.* The course website contains gene expression data from a case-control study of prostate cancer (`Singh2002`). Carry out a t-test for differential expression between cases and controls, and apply the following multiple comparison adjustment procedures with an error rate (FWER/FDR/local FDR/etc.) of 10%:

   - Bonferroni
   - Holm
   - FDR (Benjamini-Hochberg)
   - FDR with estimation of π_0
   - Local FDR (there are a variety of choices and software packages you could use; do whatever you feel is appropriate, but describe the approach you used)

   For each method, report both the number of significant results and the largest p-value that was still considered significant.

6. *Breast cancer gene expression in response to estrogen.* The data set `Scholtens2004` is from an experiment to identify genes in ER+ breast cancer cells that respond to estrogen; a more detailed description of the experimental design is available online. Analyze the data to produce three lists of genes: genes that respond to estrogen (broad response across both times), "early responders" for which the estrogen response is stronger in the short term than it is later, and "late responders" for which the estrogen response is stronger later than it is in the first 10 hours.

   For this assignment, write up a brief "Methods" section and "Results" section, as it might appear in a scientific journal, each consisting of one or possibly two paragraphs, describing what you did (Methods) and what you found (Results). For the methods section, you must use the moderated testing approach we discussed in class, where the variance estimates are borrowed across genes, but everything else is up to you – in particular, there are a variety of reasonable ways you could interpret the scientific questions and address the multiple testing issue.

   For the results section, describe the number of genes you found in each category, the FDR/significance criterion you used, and a list or table of 2 or 3 representative genes from each category so that I can check whether your results make sense.