MSCR 509: High Dimensional Analysis

Homework 9

Anish Shah

April 6, 2020

Question 1

Consider the problem of disease-associated gene identification (See attached dataset cancer.sas7bdat). The data contains 33 cancer cases and 19 normal cases (rows). The expression of 10,000 genes (columns) was measured in these samples. We want to identify genes that differentially express between cancer samples and normal samples. Complete the following:

Part A

Have SAS conduct 10,000 t-tests for a difference in average expression between cancer and normal cases for each of the 10,000 genes (you may use either method discussed in class, or a method of your own devising).

```
# Lengthen data so each experiment is a single row
df <-
    cancer %>%
    pivot_longer(-c(id, group), names_to = "gene", values_to = "expression")

# T-tests
results <-
    df %>%
    group_by(gene) %>%
    do(tidy(t.test(.$expression ~ .$group)))
```

This method of applying 10,000 tests is tedious. It was performed using R, as seen above. This resutls in 10,000 t-tests performed.

Part B

Applying no correction for multiple testing, how many genes would we declare significant at significance level=0.05? Applying the Bonferroni correction, how many genes would we declare significant at an overall alpha of 0.05? Applying the FDR correction, how many genes would we declare significant if we control the FDR to 0.05?

Using the above data, without correcting for multiple testing, out of the original 10000, there are 1010 significant genes that meet uncorrected $\alpha = 0.05$. After applying the **Bon Ferroni correction**, there are 11 significant genes. . Using teh **FDR correction**, also known as the **Benjamini-Hockberg correction**, there are 71 significant genes.

Part C

Apply the FDR correction while controlling the FDR to 5%, 10%, and 20%. How many genes would we declare significant in each case? How many false leads would we expect in each case?

With the $FDR = 5$, the number of genes that are significant would be 71. For $FDR = 10$, the number would be 107. For $FDR = 20$, the number would be 247. In each case, we would expect a proportion (the FDR level), of the significant findings to be false positives. Thus, 5%, 10%, and 20% of the significant findings.

Part D

Using the Heatmap code (cancer_heatmap.sas) provide a picture of significant genes using different methods. Comment on the results (You are NOT required to understand the programming of the code since it involves some sophisticated programming).

As seen in this heatmap, the significant findings by method of p-value adjustment is quite different. Using an uncorrected approach, there are a large number of important genes, but with strict correction by Bonferroni, we have only a handful of important genes. However, using an FDR approach finds a moderate, in-between, approach allowing us to have several potentially useful genes to explore without an excessive amount.