Problem Set 5

Stats 506, F19

Due: Wednesday December 4 by 5 pm

Instructions

- Submit the assignment by the due date via canvas. Assignments may be submitted up to 72 hours late for a 5 point reduction.
- All files read, sourced, or referred to within scripts should be assumed to be in the same working directory (`./`).
- Your code should be clearly written and it should be possible to assess it by reading it. Use appropriate variable names and comments. Your style will be graded using the style rubric (./StyleRubric.html) [10 points].
- Some of these exercises may require you to use commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Part of the purpose of these exercises is for you to learn to be resourceful and self sufficient. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.
- Except where explicitly instructed otherwise, you should do this assignment entirely in R and use the `data.table` package for working with data.
- Peer review: We will experiment with anonymous peer review for this assignment. Please complete your peer reviews by *Friday, December 6 at 5pm*. We will not revise this assignment.

Questions

Question 1 [20 points]

Repeat question 2 from problem set 4 using R and the data.table package.

Question 2 [50 points]

In this question, you will use `data.table` to clean and analyze the DNA methylation data available here (ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE138nnn/GSE138311/matrix/). You can read more about the experiment and data here (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138311)

a. Download the data and inspect the first 100 rows at the command line. Write the commands for doing so in your solution. How many lines of header information are there?

b. Use `fread` to read the data into R as a data.table. Filter the data to those probes (`ID_REF`) corresponding to chromosomal locations beginning with "ch" and discard the sample with all missing values. Finally, pivot the data to a longer format using `melt` , with each row corresponding to a single sample-probe pair.

c. Refer to the second link above to determine which samples correspond to individuals with Crohn's disease and which to non-Crohn's samples. Add a column `sample_group` to the data.table *by reference* recording this information.

d. Create a new data.table by computing a t-statistic (with homogeneous/pooled variance) comparing the difference in means between groups for each unique probe. Hint: refer to this document (https://open.umich.edu/sites/default/files/downloads/f12-stats250-bgunderson-statsfullyellowcard_0.pdf) if needed.

e. Add a column `probe_group` *by reference* assigning probes to groups using the first 5 digits of the probe ID.

f. Compute the proportion of probes within each `probe_group` that are nominally significant at the 5% level assuming a two-tailed test. Produce a figure comparing these percentages. Which group stands out as potentially over-represented?

g. Next, we will use permutation tests to assess the statistical significance of each probe group, using one of the custom statistics below. Write a function taking three arguments: (1) the data.table produced in part c, (2) a type (two-tailed, greater, or lesser), and (3) a logical flag "permute". Your function should compute t-statistics as in part (d), after, when the permute flag is true, permuting the sample group labels, and then compute the appopriate statistic from the list below. Here G is the number of probes in a given group, t^*_α is the α-quantile from a t-distribution (with appropriate degrees of freedom) and t_i is the t-statistic for probe i.

   i. "two-tailed": $T_{abs} = \frac{1}{G}\sum_{i=1}^{G} |t_i| 1[|t_i| > t^*_{1-\alpha/2}]$
   ii. "greater": $T_{up} = \frac{1}{G}\sum_{i=1}^{G} t_i 1[t_i > t^*_{1-\alpha}]$
   iii. "lesser": $T_{down} = \frac{1}{G}\sum_{i=1}^{G} t_i 1[t_i < t^*_{\alpha}]$

h. Use your function to compute the T_{abs} score for each probe group on the original data. Then, use your function to compute scores for each of 1,000 permutations and compute p-values for testing whether the observed T_{abs} score for each group is larger than expected under the null hypothesis that patterns of gene expression are the same across the Crohn's and non-Crohn's groups. Time how long it takes to compute the 1,000 permutations.

i. Repeat the previous part using the T_{up} score and using `mclapply` for parallelism.

j. Repat the previous part using the T_{down} score and using futures for parallelism.