MATH 185 – Homework 4
Due Monday, 05/16/2016, by 11:59 PM

Send your code to `math185ucsd@gmail.com`*. Follow the following format exactly. For Homework 1, in subject line write "MATH 185 (HW 1)" and nothing else in the body. There should only be one file attached, named* `hw1-lastname-firstname.R`*. Make sure your code is clean, commented and running. Keep your code simple, using packages only if really necessary. If your code does not run, include an explanation of what is going on.*

Problem 1. (Two-way ANOVA by permutation) Read the slides on the two-way layout (starting after the introduction of the Friedman test). Write a function `twowayPermTest(dat, B = 999)` which takes a data frame `dat` with a two-way layout and returns a p-value for the treatment sum of squares calibrated by permutation (within each block). Apply your function to the `ToothGrowth` dataset (loaded by default).

Problem 2. (Microarray data: multiple t-tests) Microarrays enable the biologist to measure the expression level of genes, in humans or other species. In some studies, the gene expression level is measured in a number of individuals, some with a disease and some without (control group), to identify what genes are differentially expressed. The goal is to understand the role of genes in the development of the considered disease. Read as much of the following paper as you feel inspired to, and enough to understand the basics: http://www.pnas.org/content/96/12/6745.full
The corresponding dataset (after some processing) is available here:
https://github.com/ramhiser/datamicroarray/blob/master/data/alon.RData
with a brief explanation available here:
https://github.com/ramhiser/datamicroarray/wiki/Alon-%281999%29
A simple (and somewhat outdated) method for identifying differentially expressed genes is to perform a (two-sided) Student t-test for each gene comparing the two groups. Write code that carries this out. The result should be a p-value for each gene. Then apply all the methods seen in lecture for controlling the FWER or the FDR. Offer some brief comments.

Problem 3. (Microarray data: multiple permutation t-tests) With the same dataset, perform a permutation t-test for each gene. (In other words, calibrate the t-ratio statistic by permutation.) Doing this might be safer given the fact that the group sizes are not very large. One has the choice of permuting the subjects for each gene independently, or permuting the subjects and recomputing all the t-ratios for all the genes. (Of course, either way, the process is repeated many times.) Implement the latter, which is faster (and is preferred for other reasons).

Remark. In Problems 2 and 3, it is reasonable to presuppose that there is some dependency between the gene expression levels in each individual. We saw in lecture that some methods can deal with any dependency (e.g., Bonferroni for FWER control). However, one can gain in power by "adapting" to the underlying dependency. See, for example, the permutation-based method proposed in (Westfall and Young, 1993), for FWER control. See also (Romano and Wolf, 2005).