BIODS215-2018 Problem set 1

Due date: 2018/1/30, Release date: 2018/1/18

Yosuke Tanigawa, Manuel Rivas, James Zou, and Julia Salzman

- Please write the answer to the problem set in one pdf document that includes all the codes and results from computational experiments.
    - Jupyter Notebook (https://http://jupyter.org) and R Markdown (http://rmarkdown.rstudio.com) have functionalities to export a document into a pdf file.
- Please submit your answer through gradescope (https://gradescope.com , Entry Code:MWYKE4).

1. Introduction to Biomedical Data Science (20 pts)

a. Briefly describe why data science is becoming more important for biomedical applications using some examples.

b. What is a difference between inference and prediction? Please describe it briefly and give an example for each of the inference problems and prediction problems in the context of biomedicine.

c. Describe situations where people should and should not use deep learning in the context of biomedicine.

2. Robust permutation tests (80 pts)

- We've learned permutation tests in the class. The following questions are based on the biological examples we've covered in the class and the results from E. Chung & J. Romano, 2013 (doi:10.1214/13-AOS1090) and aiming you to teach you a way to construct a robust way of a permutation test.

a. In what situation, is it difficult to construct a level test based on permutation tests?

b. Describe a biomedical example where people are commonly using permutation tests with possibly inflated rejection probabilities.

- Let's say your experimental collaborator is interested in the effect of drug *D* on the expression level of a particular gene *G*. They collected the expression level of gene *G* in two conditions: treated and control. For each condition, they measured the gene expression level of the gene *G* for *k* cells though single-cell RNA-seq experiment and handed the data to you. Your goal as a data scientist is (1) to estimate the expression levels of gene *G* from the read counts and (2) to compare the mean of the expression levels between two groups and assess the significance of the difference if any. Given the noisy nature of single-cell experiments, you performed the quality assessment analysis and found that you have data with sufficient quality for *n* and *m* cells ($n < k, m < k$) for treated and control, respectively (we assume there is no censoring issue in this step, i.e. data will come with bad quality at random).

c. What is the probability distribution commonly used to model the expression level of genes based on read counts?
d. Describe a procedure to assess the significance of the difference between the observed mean of expression levels using permutation test. Note that even the two group has the same average expression levels, you may observe a difference of means due to the noise of experiment.
e. Describe a studentization procedure for the distribution you've answered in the previous question (question 2c).
f. Using any programming language of your choice, implement a Monte Carlo simulation for permutation tests for Poisson distributions. We will ask you to include your code in your answer. This question is meant to give you a partial credit in case you are not able to fully answer the following question.
g. Perform a numerical experiment to compare the robust permutation methods introduced by E. Chung & J. Romano, 2013 and the permutation without studentization. You will generate a data under the null (the two groups have the same expression level) and try to see how many times each of the methods will reject the null hypothesis. Try several different choices of *(m, n)* and see how the results may change with respect to these parameters for each of the permutation method. You may find it useful to check chapter 4 and table 1 of the E. Chung & J. Romano, 2013 paper: we are basically asking you to perform a similar analysis with a different probability distribution.