MATH 185 – Homework 3
Due Monday, 05/02/2016, 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. (Approximate permutation tests) In a two-sample situation, tests are typically calibrated by permutation. However, for the P-value to be strictly valid, this requires that the two samples come from the same population (i.e., have the same distribution) under the null hypothesis. What happens when this is not the case? To explore this situation, consider testing about the means and consider the permutation test based on the difference in sample means. We place ourselves in a setting where $X_1, \dots, X_m \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \dots, Y_n \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, all observations assumed independent. *We place ourselves under the null hypothesis* and examine how the test statistic behaves. We choose $\mu_X = \mu_Y = 0$ (without loss of generality) and[1] $\sigma_X^2 = 1$ and $\sigma_Y^2 = 5$. In that case $\bar{X} - \bar{Y} \sim \mathcal{N}(0, \frac{1}{m} + \frac{5}{n})$. How does it compare with the distribution of the statistic after a random permutation of the data? (Ideally, these distributions would be the same, but here the two samples do not come from the same population, so we cannot expect this to be strictly true.)

A. Assume that $m = 1{,}000$ and $n = 3{,}000$. Generate two samples as above and then permute them once to obtain one value of the difference in means after permutation. Repeat the whole thing 10,000 times. How does the distribution (here sampled by Monte Carlo) compare with the distribution before permutation (meaning $\mathcal{N}(0, \frac{1}{m} + \frac{5}{n})$)?

B. Repeat with $m = n = 2{,}000$.

Problem 2. (Tests based on runs) In a two-sample setting, with an X sample and a Y sample, an X- run is sequence of consecutive X's in the ordered combined sample. For example, if the pattern of the ordered combined sample is $XXYXXXYXYY$ (so that the sample sizes are 6 and 4, respectively) there are 3 X-runs of respective lengths 2, 3, and 1. The *number of runs test* is based on the number of X-runs (proposed by Wald and Wolfowitz in 1940). The *longest run test* is based on the length of the longest X-run (proposed by Mosteller in 1941). Let's focus on the former.

A. What is the most appropriate null hypothesis for such a test?

B. Find a function in R (possibly in some package) that implements that test. Apply it to the `cloud seeding` dataset.

C. Write your own function `nb.runs.test(x, y, B=999)` that implements the two-sided version of that test. Say in a few words how you compute the p-value. Is the p-value exact?

D. How would a one-sided version of the test look like?

Problem 3. (Test for symmetry) Typically, a two-sample numerical test can be turned into a (one-sample) test for symmetry (about 0 by default) by simply considering the positive and negative parts of the sample as different samples. Think about how you would do that starting from the number of runs test. Then write a function `nb.runs.sym.test(x, B=999)` that does exactly that. Test your function on some synthetic data.

[1] The variances need to be different, otherwise the two samples come from the same population and we know that things go well, meaning that the distribution of the test statistic after a random permutation of the data remains unchanged.