MATH 185 – Final

Due Wednesday, 06/13/2017, by 11:59 PM

Send your code to math185ucsd@gmail.com*. Follow the following format exactly. In subject line write "MATH 185 (Final)" and nothing else in the body. There should only be one file attached to the email, named* final-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 (as a comment in the code) of what is going on.*

AGREEMENT

By taking this exam, you agree to not discuss the exam with anyone, starting now, neither with a classmate or anyone else, neither in person nor through other means, including electronic. Unless otherwise specified, it is acceptable to copy-paste from the lecture or homework solution code.

Problem 1. (A simple example where the bootstrap fails.) Consider the situation where $X_1, \dots, X_n$ are iid from a distribution with mean θ and variance 1. It is desired to provide a confidence interval for $|\theta|$. The plug-in estimator is $|\bar{X}_n|$, the absolute value of the sample mean. The pivotal bootstrap confidence interval is based on estimating the distribution of $|\bar{X}_n| - |\theta|$ by the bootstrap distribution of $|\bar{X}_n^*| - |\bar{X}_n|$. This happens to fail when $\theta = 0$. (This is because the absolute value, as a function, is not smooth at the origin.) We examine the situation when $X_1, \dots, X_n$ are iid standard normal.

A. Compute the distribution function of $\sqrt{n}(|\bar{X}_n| - |\theta|) = \sqrt{n}|\bar{X}_n|$ in closed form (no need to show your work) and draw it. (It does not depend on n.)

B. Generate a sample of size $n = 10^6$ and estimate the bootstrap distribution of $\sqrt{n}(|\bar{X}_n^*| - |\bar{X}_n|)$ using $B = 10^4$ replicates. Add the resulting empirical distribution function to the plot.

C. Offer some brief comments.

Problem 2. (Meta-analysis) In meta-analysis, the goal is to gather information, and perform inference, based on several studies published by different researchers, over several years. For example, consider 8 studies[1] conducted between 1981 and 1984, comparing an experimental surgical intervention (proximal gastric vagotomy) to an established intervention (truncal vagotomy plus drainage). Here the event of interest is recurrence, the control group is truncal vagotomy plus drainage, and the treatment group is the experimental intervention.

Study	Treatment	Control
1	9/48	6/50
2	15/70	5/67
3	1/68	4/69
4	5/59	6/64
5	7/71	2/74
6	8/37	6/37
7	13/76	8/75
8	9/56	4/50

It reads as follows. Take Study 1, corresponding to the first row. There were 48 individuals in the treatment group, 9 of which had a recurrence, and there were 50 individuals in the control group, 6 of which had a recurrence. The results of each study are often summarized in a 2-by-2 contingency table. The goal is to assess whether the new procedure yields a smaller rate of recurrence than the standard procedure.

A. As a preliminary, apply the (one-sided) Fisher exact test to each of these 8 studies, obtaining 8 p-values. Store these in a vector called pval.

B. Apply the (one-sided) *Kolmogorov-Smirnov test* to the p-values. What is the null distribution in the present context? Explain.

C. Apply the *Liptak-Stouffer test* to the p-values. This test rejects for large values of $T = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \Phi^{-1}(1 - P_j)$, where Φ is the standard normal distribution function and P_j is the p-value associated with the j-th study. Note that, when $P_1, \dots, P_m$ are IID uniform in $[0, 1]$, T has the standard normal distribution. Implement the function yourself and name it ls.test.

D. Apply the *Cochran-Mantel-Haenszel test* (directly to the data, without going through the p-values). You can read about the test here, although the formula is more rigorously written here. No need to implement the test yourself.

[1] Buyse, M., P. Hewitt, and M. Koch. Data analysis for clinical medicine: the quantitative approach to patient care in gastroenterology. International University Press, 1988.

Problem 3. (Selecting the degree of a polynomial model) Write a function poly.fit(x, y, stop = 0.05) that takes in the predictor and response variable vectors, and fits a polynomial model where the degree is chosen sequentially, starting at degree = 0, increasing the degree by 1, and stopping when the improvement in R-squared is less than specified by stop. The function should return the least squares coefficient of the final polynomial model. Apply your function to the dataset steam in the MASS package.